UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
CADERNOS DE MATEMÁTICA E ESTÁTISTICA
SÉRIE B: TRABALHO DE APOIO DIDÁTICO

ESTATÍSTICA DESCRITIVA II

DINARA WESTPHALEN XAVIER FERNANDEZ

SÉRIE B, Nº 24
PORTO ALEGRE, AGOSTO DE 1994

APRESENTAÇÃO

O material de ESTATÍSTICA DESCRITIVA II foi organizado a partir de nossa experiência como docente nessa disciplina no Curso de Bacharelado em Estatística da UFRGS. Para que o conteúdo da referida disciplina seja totalmente coberto, é necessário consultar a publicação NÚMEROS ÍNDICES também de nossa autoria.

Quando trabalhamos com essa disciplina, introduzimos o uso do software estatístico MINITAB, realizando aulas práticas no Laboratório de Recursos Computacionais, do Instituto de Matemática da UFRGS. Anexamos algumas dessas aulas, a título de sugestão.

Também consta na Bibliografia, vários artigos que foram alvo de Seminários apresentados pelos alunos ao final do semestre. Desta forma, já no início do Curso, o aluno é incentivado a leitura de bibliografia estrangeira, bem como oferecemos a oportunidade de desenvolver sua habilidade de expor publicamente.

Essas notas tem como principal objetivo facilitar o desenvolvimento de disciplinas introdutórias de Estatística, mas não pretende substituir a consulta e o enriquecimento com outras bibliografias.

A autora.

ÍNDICE

I- NÍVEIS DE MENSURAÇÃO

1. CARACTERIZAÇÃO

Mensuração é o processo que consiste em atribuir números a objetos, indivíduos ou eventos.

O grau de mensuração que se atinge é função das regras sob as quais se fez a atribuição dos números. Quanto mais tipos de comparações forem permitidas por uma classificação com relação a uma variável, tanto mais elevado é o seu nível de mensuração. Quando classificamos os alunos de uma escola de acordo com as variáveis "nacionalidade" e "rendimento escolar", podemos, em relação à primeira variável, apenas dizer se dois alunos tem a mesma nacionalidade ou não; já em relação à outra variável, podemos dizer se os alunos tem o mesmo rendimento escolar ou não e ainda se um aluno tem rendimento maior ou menor que o outro. Assim, a variável "rendimento escolar" atingiu um nível de mensuração mais elevado que a variável "nacionalidade".

As operações e relações utilizadas na obtenção dos conjuntos de valores definem e limitam as operações possíveis que devem estar associadas à estrutura numérica à qual a mensuração é isomorfa.

Estudaremos quatro níveis de mensuração: **nominal**, **ordinal**, **intervalar** e escala **de razão**.

As escalas **nominal** e **ordinal** são os tipos mais comumente encontrados nas ciências do comportamento e os dados mensurados por essas escalas devem ser analisados por métodos não-paramétricos. As variáveis qualitativas são classificadas em ordinais, quando houver um sentido de ordenação entre seus valores e em nominais , quando não houver sentido de ordenação entre seus valores, como é o caso da variável "nacionalidade".

Dados mensurados em escalas **intervalar** ou **de razão** podem ser analisados por métodos paramétricos, desde que as suposições do modelo estatístico sejam satisfeitas.

São quatro os tipos de comparação que uma variável de nível de mensuração numa escala **de razão** nos permitem fazer:

1ª) Identificar se objetos, indivíduos ou eventos pertencem ou não a uma classe da variável.

Dados dois objetos, indivíduos ou eventos pertencentes a classes diferentes, identificar, em relação à variável em questão, se:

2ª) a classe a que um pertence é maior ou menor do que a classe a que outro pertence.

3ª) a diferença entre as classes a que dois deles pertencem é maior, igual ou menor que a diferença entre as classes a que outros dois pertencem.

4ª) um deles é, por exemplo, o dobro do outro. Isto é, a razão entre dois valores quaisquer da escala é conhecida. Isto ocorre porque existe um zero absoluto.

Se a última comparação não pode ser realizada, o nível da escala é **intervalar**. Se as duas primeiras comparações podem ser feitas, temos uma escala **ordinal** e se apenas a primeira é válida, estamos trabalhando com uma escala **nominal**.

Consultar SIEGEL,S., como complementação.

2. EXERCÍCIOS

1. Guiando-se pelas comparações listadas acima, preencha os espaços em branco.

Tomemos a classificação das escolas em relação à variável "número de alunos". As classes a que as escolas podem pertencer são: ...,50,51,...600,...

______________________dizer se uma escola pertence ou não
(podemos/não podemos)

pertence a classe de escolas que possuem 350 alunos. Portanto, a classificação de escolas em relação à variável ______________ estabelecer o primeiro tipo de
(permite/ não permite)
comparação.

Dadas duas escolas, por exemplo, de 400 e 320 alunos ______________ dizer se a classe "número de alunos"
(podemos/ não podemos)
a que uma pertence é maior ou menor do que a classe a que outra pertence. Portanto, a classificação com relação à variável ______________ estabelecer o segundo tipo de
(permite/não permite)
comparação.

Dadas quatro escolas, por exemplo, com 500, 1000, 700 e 240 alunos, ______________ dizer se a diferença entre as
(podemos/não podemos)
classes a que duas delas pertencem é maior, menor ou igual à diferença entre as classes a que as outras duas pertencem. Portanto a classificação em relação à variável ______________ estabelecer o terceiro tipo de
(permite/não permite)
comparação.

Dadas duas escolas uma com 500 e outra com 1000 alunos,

______________ dizer que a segunda possui o dobro de
(podemos/não podemos)
alunos que a primeira. Portanto, a classificação ______________ estabelecer o quarto tipo de
(permite/não permite)
comparação.

A classificação com relação à variável "número de alunos" nos permite estabelecer os quatro tipos de comparação, portanto, é uma variável de nível de mensuração ______________

2. Analise as comparações possíveis relativamente às variáveis:

2.1. "temperatura média no inverno" de regiões do globo terrestre. As classes em que uma região do globo terrestre podem ser classificadas são: ..., $-5^{o}C$, ..., $0^{o}C$, ..., $12^{o}C$, ...

2.2. "nível sócio-econômico" dos alunos de uma escola obtido a partir de informações de um questionário. Definimos cinco classes: nível 2(alto);nível 1(médio);nível 0(médio); nível -1(médio baixo) e nível -2 (baixo).

2.3. "conceito" em Ingles dos alunos de uma Escola. As classes a que um aluno pode pertencer são: A (Ótimo), B (Bom), C (Médio), D (Regular) e E (Fraco).

3. Identifique o nível de mensuração das variáveis abaixo definidas e justifique:

3.1. "número de estabelecimentos de ensino sob a jurisdição de uma Delegacia" segundo as quais se classificam Delegacias de Ensino de um dado sistema com as classes: 0-10, 11-20,...

3.2. "tipo de estabelecimento de ensino" segundo a qual as escolas pertencem a uma das classes: só 1º grau, só 2º grau, 1º e 2º grau.

3.3. "assuidade do professor" segundo a qual se classificam os professores quanto ao número de faltas obtidas num mes, nas classes: nunca falta, falta muito pouco, falta razoavelmente, falta muito.

4. Um entrevistador classifica a reação do indivíduo quanto a direção do olhar pelos números 0, 1, 2 onde 0 representa olhar para baixo; 1 representa olhar para cima e 2 representa olhar diretamente para o entrevistador. Qual é a escala de medida adotada ?

5. Um sujeito atribui a um grupo de produtos alimentares uma "graduação" por ordem de "gosto": grau 1, ao alimento menos gostoso de todos e grau 10 ao mais gostoso. Poderíamos afirmar que a diferença em gosto entre os alimentos de grau 3 e 4 é equivalente à diferença em gosto entre os itens 7 e 8 ? Justifique e diga qual é o nível de mensuração desta variável.

6. A variável psicológica "prestígio" pode ser medida em função de unidades estabelecidas e padronizadas, análogas a graus (temperatura) ou segundos (tempo) ? Justifique.

7. Qual é o nível de mensuração da variável "tempo de reação de um indivíduo a determinado estímulo" ?

8. Ao medir a "pressão na tecla de resposta", estamos utilizando uma escala ______________________ pois podemos afirmar que uma pressão de 3 libras é tres vezes superior a uma de 1 libra e, além disso, zero libra representa a inexistência de pressão.

9. Realize o Estudo Dirigido sobre Níveis de Medição que segue.

ESTUDO DIRIGIDO SOBRE NÍVEIS DE MEDIÇÃO

INTRODUÇÃO: A finalidade da medida nas ciências do comportamento é a de oferecer bases exatas, objetivas e comunicáveis para descrever, diferenciar e classificar as características e os comportamentos dos indivíduos.

O nível menos refinado de medida, a **MEDIDA NOMINAL** é a classificação das características ou dos comportamentos tendo por objetivo a simples identificação. Um nível um pouco mais refinado de medida permite a ordenação das categorias em termos do grau ou quantidade do fenômeno que está sendo medido; a isto se chama **MEDIDA ORDINAL**. O nível seguinte é a **MEDIDA INTERVALAR**, no qual a distância entre duas categorias quaisquer é conhecida. O nível mais refinado de medida é alcançado quando a escala a medir o fenómeno tem um verdadeiro ponto zero em sua origem. Tal escala é chamada de **ESCALA DE RAZÃO.**

Neste estudo dirigido, você irá tomar conhecimento das limitações de cada tipo de medida e da espécie de fenômeno para o qual cada tipo é aplicável.

ATENÇÃO: Dobre a folha à margem direita. Leia atentamente o primeiro ítem, responda e depois confira sua resposta. Só passe para o ítem seguinte após ter compreendido bem o anterior.

	dobre aqui
1. Dar apenas nomes ou rótulos a um conjunto de objetos, de carecterísticas ou de comportamento é o nível ____________ (mais/menos) refinado de medida.	menos
2. O termo medida nominal é usado para designar o tipo de medida considerado quando nós apenas ____________________ (rotulamos/damos nomes) às categorias compreendidas numa característica.	rot/ d.n.
3. Quando damos nomes que estabelecem diferenças entre uma uma categoria e outra, mas não estabelecemos nenhuma ordem entre as categorias, estamos usando o tipo de medida _____	nom.
4. Por exemplo: se desejássemos designar as duas categorias compreendidas na característica "sexo", as palavras "masculino" e ______________ seriam usadas.	fem.
5. As palavras "masculino" e "feminino" são rótulos que usamos para designar as duas categorias compreendidas na característica "sexo". Esses dois nomes de categorias _______________ (implicam/não implicam) em que uma categoria seja "melhor" do que a outra.	n.i.
6. As categorias "masculino" e "feminino" não tem uma ordem implícita; seu uso, portanto, constitui uma medida ________	nom.
7. Se categorizarmos as preferências políticas em "democrata" e "republicano", a ordem __________(tem/não tem) importância.	n.t.

8. Essas categorias não admitem ordem; portanto, elas ________ (tem/ não tem) valor quantitativo.	n.t.
9. Os nomes "democrata" e "republicano" são categorias de preferência política que não possuem ordem implícita; portanto, o uso dessas palavras constitui uma medida ________	nom.
10. Medida nominal, pois, é o processo de diferenciar comportamentos ou características e de dar ________ (nomes/rótulos) às categorias não-ordenáveis correspondentes a esses comportamentos ou características.	n/r
11. As categorias "protestante", "católico" e "judeu" ____ (são/não são) ordenadas.	n.s.
12. Portanto, o uso das categorias "protestante", "católico" e "judeu" constitui uma medida ________	nom.
13. As categorias "excelente", "bom" e "regular" ________ (são/não são) ordenadas.	são
14. Portanto, as categorias do item 13 ________ (representam/não representam) uma medida nominal.	n.r.
15. As categorias "concorda plenamente", "concorda" e "discorda" são ordenadas e permitem uma medida ________(mais/menos) refinada do que as categorias nominais.	mais
16. Se tivéssemos que distinguir apenas os masculinos dos femininos, os "anões masculinos" e os "gigantes masculinos" ________(recairiam/não recairiam) na mesma categoria.	rec.
17. Entretanto, se as categorias "abaixo da média" e "acima da média", quanto à altura, fossem dadas como classificação então os "gigantes masculinos" e os anões masculinos" ____ ________(recairiam/não recairiam) na mesma categoria.	n.r.
18. Nos exemplos anteriores, as categorias internas à "sexo" ________(estão/não estão) ordenadas; as categorias internas à variável "altura" ________(estão/não estão) ordenadas.	n.e. e.
19. No exemplo do "anão-gigante", "sexo" ________ (representa/não representa) uma medida nominal; "altura" ________(representa/não representa) uma medida nominal.	r. n.r.
20. As categorias usadas na medida nominal ________ (tem/não tem) relação especial uma com a outra.	n.t.
21. Entretanto, se as categorias estiverem ordenadas, é claro que ________(há/não há) uma relação entre elas.	há

22. As categorias "concorda plenamente","concorda" e "discorda" especificam uma relação ordenada quanto ao grau de ______________ (concordância/discordância)	c.
23. A característica "grau de concordância" está enunciada numa escala ordinal de medida porque as categorias permitem uma ____________ qualitativa.	orde.
24. Dividir a característica "situação sócio-econômica" em tres categorias: "baixa", "média" e "alta" é um exemplo do uso da escala de medida ____________.	ordi.
25. Se tivéssemos que atribuir números às tres categorias de situação sócio-econômica, "baixa" poderia ser designada pelo número "1".Para evidenciar uma certa ordem,poderíamos atribuir o número ___ à "média" e o número ____ à "alta".	2 3
26. Em alguns casos especiais usamos letras como símbolos para designar a ordenação. Os conceitos ou graus escolares são o exemplo mais típico: caso em que "A" é mais alto do que "__" e este mais alto do que "__".	B C
27. A direção ordenada com relação a "pouco importante", "relativamente importante" e "muito importante"__________ (está/não está) claramente indicada.	está
28. Embora a direção esteja indicada,a distância ou intervalo entre as categorias consideradas _________ (está/não está) indicado.	n.e.
29. Através de uma escala classsificatória como essa poderíamos saber que "muito" é mais importante do que "relativamente" e que "relativamente" é ____________ (mais/menos) importante do que "pouco".	mais
30. A distância entre "muito importante" e "relativamente importante" provavelmente ____ (é/não é) a mesma do que a distância entre "relativamente importante" e "pouco importante".	n.é
31. A ordem classificatória portanto não implica em qual é a _________________ entre as categorias.	dist.
32. A ordenação das características ou dos componentes estão colocados numa escala de medida ___________.	ordi.
33. As características e os comportamentos que não podem ser ordenados exigem o tipo de medida ____________.	nomi
34. Uma escala ordinal __________ (implica/não implica) em direção enquanto uma escala nominal ____________ (implica/não implica) em direção.	i. n.i.
35. As escalas ordinais oferecem um grau ________ (menos/mais) preciso de medida do que as escals nominais.	mais

36. Embora uma escala ordinal apresente direção,as distâncias entre as categorias ou os pontos de escala são ______ (iguais/desiguais).	d.
37. Obteríamos mais precisão na medida,se além da direção, a escala tivesse distâncias ______(iguais/desiguais) entre os pontos que possui.	i.
38.As distâncias entre as categorias ou pontos numa escala são chamadas de intervalos.Assim, uma escala que tenha uma distância fixa entre os pontos é chamada de escala de ____ (intervalos/pontos).	int.
39.Considere a distância entre as temperaturas 20°C e 30°C e entre 50°C e 60°C. As distâncias são ambas de _____ $^\circ$C e são, portanto, __________ (iguais/desiguais).	10 i.
40. Há uma unidade fixa de ________ na escala de temperatura.	med.
41.As distâncias entre os pontos numa escala de temperatura são _________.	i.
42. A temperatura é um exemplo de medida de ____________.	int.
43. As distâncias entre os pontos numa escala ordinal são _______ (iguais/desiguais). Já numa escala de intervalos elas são ________.	d. i.
44. Por exemplo:o intervalo entre "regular" e "bom" provavelmente ___ (é/não é) igual ao intervalo entre "bom" e "excelente".	n.é
45. A atribuição dos números "1" para "regular", "2" para "bom" e "3" para "excelente" ________ (iguala/não iguala) os intervalos entre as categorias.	n.i.
46. Atribuir números às categorias ordenadas não faz com que os intervalos entre elas se tornem _________ .	i.
47. Os números atribuidos aos pontos de uma escala ordinal _____(são/ não são) aditivos, porque as distâncias entre os pontos da escala ______ (são/não são) iguais.	n.s. n.s.
48. Com efeito,muitas operações aritméticas podem ser efetuadas em escala de __________, mas não em escala _______ ou em escala ___________.	int. ord. nom.
49.Vamos supor que cinco indivíduos façam o teste de soletração composto por palavras aproximadamente iguais quanto ao nível de dificuldade, e que os escores sejam colocados numa escala de intervalos. Seus escores de 25,22, 19, 13 e 11 _____________ (poderiam/não poderiam) ser somados.	p.

50.Numa escala de intervalos os escores podem ser subtraí-dos, multiplicados e divididos porque a escala tem ______ iguais.	int.
51. Já que se pode realizar uma grande variedade de operações aritméticas com escals de __________, elas são mais úteis do que as escalas de ______________ e __________.	int. ord. nom.
52. Comparadas às escals ordinais, as escalas de intervalo são escalas de medida _____ (mais/menos) refinada.	mais
53.Para a maioria dos testes elaborados pelos professores, os intervalos entre os escores brutos não são _________.	ig.
54. Nos testes educacionais e psicológicos rigorosamente elaborados os _________ serão sempre iguais.	int.
55. Assim,para testes padronizados,os escores são frequentemente tratados como se estivessem numa escala de medida _____________.	int.
56. São exemplos disso, a classificação por grupo e os resultados de QI.A maioria das pesquisas realiza muitas operações aritméticas com tais escores,tratando-se, portanto, como se _________ (estivessem/não estivessem) numa escala intervalar.	e.
57. O quarto e mais refinado tipo de medida é a escala de razão.Esta escala possui o que se chama ponto zero absoluto, que é o ponto no qual ____ (há/não há) escore.	n.há
58. Há escalas de razão nas ciências físicas porque as variáveis tais como altura, peso e tempo tem um ponto ______ absoluto.	zero
59. Já que a variável "peso" tem um ponto zero absoluto ___ (é/não é) correto concluir que um indivíduo pesando 90 quilos é duas vezes mais pesado do que um outro que pese 45 quilos.	é
60. Considerar que um indivíduo que tem QI 100 seja duas vezes mais inteligente do que um indivíduo de QI 50 é______ (certo/errado).	e.
61. Considerar que um QI de 100 indique o dobro de inteligência de um QI de 50 seria considerar incorretamente que a variável "inteligência" tem o ponto ______ absoluto.	zero
62. Para os testes educacionais e psicológicos,o ponto zero é absolutamente arbitrário;assim,tais testes não oferecem resultados numa escala _______________ .	de r.
63.Já que não é possível estabelecer um ponto zero absoluto na medida de variáveis na ciência do comportamento, escalas _____________ não existem nesta área.	de r.

II- CARACTERÍSTICAS DE TENDÊNCIA CENTRAL E DE VARIABILIDADE

1. CARACTERÍSTICAS DE TENDÊNCIA CENTRAL

As medidas de posição ou tendência central são aquelas que tendem a se localizar em um valor central dentro de um conjunto de dados.

1.1 MÉDIAS

1.1.1. MÉDIA ARITMÉTICA

Notação: μ (média populacional)

$\overline{X}$ (média amostral)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{(dados não agrupados)}$$

$$\mu = \frac{\sum_{i=1}^{m} f_i \; x_i}{N} \quad \text{(dados agrupados)} \quad \text{onde} \quad N = \sum_{i=1}^{m} f_i$$

OBSERVAÇÃO: No caso de média amostral, o denominador é o tamanho n da amostra.

EXEMPLO :

a) QI de 12 indivíduos : 93 94 95 97 98 98 100 101 101 104 105 108

$T = \sum^{12} x_i = 1194 \quad \Rightarrow \quad \mu = 1194/12 = 99,5$

b) Organizando os dados numa distribuição de frequência, obtemos:

Tabela 1: QI de 12 indivíduos

QI	f_i	x_i	$f_i\ x_i$
90 |— 94	1		
94 |— 98	3		
98 |— 102	5		
102 |— 106	2		
106 |— 110	1		
Σ	12		

μ = 1196/12 = 99,7

OBSERVAÇÃO: Se em lugar de f_i consideramos "pesos", temos uma média aritmética ponderada, como também é chamada a média para dados agrupados.

EXEMPLO :

Produto	Preços	Quantidade
A	100	30
B	110	72
C	120	85
D	150	53

$$\mu = \frac{100.30+110.72+120.85+150.53}{30 + 72 + 85 + 53} = \frac{29070}{240} = 121,12$$

Erro relativo máximo cometido no cálculo da média para dados agrupados em distribuição de frequência

$$\varepsilon = \frac{h}{2\ \mu - h}$$

onde h = amplitude do intervalo
μ = média da distribuição de frequência

No exemplo 1:

$$\varepsilon = \frac{4}{2 \times 99,7 - 4} = 0,0205 \quad \text{, isto é,}$$

um erro máximo de 2,05 % para mais ou para menos.

PROPRIEDADES:

P1: A soma algébrica da diferença entre cada valor observado e a média aritmética é nula, isto é, a soma dos desvios em torno da média é zero:

$$\sum_{i=1}^{N} d_i = \sum_{i=1}^{N} (x_i - \mu) = 0$$

Demonstração: $$\sum_{i=1}^{N} (x_i - \mu) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu =$$

$$\sum_{i=1}^{N} x_i - N\mu = N\mu - N\mu = 0$$

OBS: Se os dados estiverem agrupados numa distribuição de frequência :

$$\sum_{i=1}^{N} f_i d_i = \sum_{i=1}^{N} f_i (x_i - \mu) = 0$$

P2: A média de uma constante é a própria constante.

P3: A média do produto de uma constante por uma variável é igual ao produto da constante pela média da variável.

P4: A média da soma de uma constante com uma variável é igual a soma da constante com a média da variável.

P5: A soma dos quadrados dos desvios da média aritmética é mínima em relação a soma dos quadrados dos desvios relativamente a qualquer outro valor distinto da média aritmética, isto é:

$$\sum_{i=1}^{N} (x_i - \mu)^2 = 0 \quad < \quad \sum_{i=1}^{N} (x_i - a)^2 = 0 \, , \quad \mu \neq a$$

Demonstração:

$$S = \sum_{i=1}^{N} (x_i - \mu)^2 = \sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + N\mu^2 =$$

$$= \sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2 \Rightarrow S = \sum_{i=1}^{N} x_i^2 - N\mu^2$$

$$R = \sum_{i=1}^{N} (x_i - a)^2 = \sum_{i=}^{} x_i^2 - 2aN\mu + Na^2$$

Fazendo R − S temos:

$$R - S = N a^2 - 2 a N \mu + N \mu^2 = N (a^2 - 2 a \mu + \mu^2) =$$
$$= N (a - \mu)^2 .$$

Como N e $(a - \mu)^2$ são positivos,

$$N (a - \mu)^2 > 0 \Rightarrow R - S > 0 \Rightarrow R > S .$$

P6: A média aritmética de um conjunto constituido por k séries é igual ao quociente da soma dos produtos das médias parciais pelos respectivos números de ocorrências sobre o número total de componentes das séries (média das médias ou média composta)

Demonstração:

$$\mu_1 = \frac{\Sigma x_i}{N_1} \qquad \mu_2 = \frac{\Sigma y_i}{N_2} \qquad \ldots \qquad \mu_k = \frac{\Sigma z_i}{N_k}$$

$$\Rightarrow \Sigma x_i = \mu_1 . N_1 \qquad \Sigma y_i = \mu_2 . N_2 \qquad \ldots \qquad \Sigma z_i = \mu_k . N_k$$

$$\Rightarrow \mu_a = \frac{\Sigma x_i + \Sigma y_i + \ldots + \Sigma z_i}{N_1 + N_2 + \ldots + N_k} = \frac{\sum^{k} \mu_i N_i}{\sum^{k} N_i}$$

EXEMPLO : Certa Agremiação Esportiva possui equipes de vôlei, basquete e futebol. O número de componentes das mesmas e as médias de alturas são:

Vôlei: 18 atletas - estatura média = 1,72 m

Basquete: 14 atletas - estatura média = 1,80 m

Futebo: 35 atletas - estatura média = 1,70 m.

Qual é a altura média dos atletas da Agremiação ?

$$\mu_a = \frac{1,72 \times 18 + 1,80 \times 14 + 1,70 \times 35}{18 + 14 + 35} = 1,726 \text{ m}$$

VANTAGENS:

1) É a medida de tendência central mais conhecida e empregada.

2) É facilmente calculável.

3) Pode ser tratada algèbricamente (propriedades).

4) Serve para comparar conjuntos semelhantes.

5) É particularmnente indicada para séries que possuem valores em progressão aritmética ou simétricos em relação a um valor médio e máximo.

6) Depende de todos os valores da série.

7) Será representativa quanto maior o número de termos.

DESVANTAGENS:

1) Não representa bem os conjuntos que revelam tendências extremas. Por exemplo: (a) X: n° de acidentes semanais

$X = \{ 2, 4, 4, 3, 2, 40 \}$ $\quad \mu_x = 9$

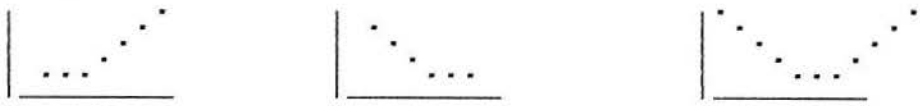
(b) Salários:

x_i	1	10	100
f_i	50	2	1

$\mu_x = 3,2$

É grandemente influenciada pelos valores extremos da série, razão porque às vezes pode não ser típica.

Seu emprego é desaconselhável para distribuições do tipo:



2) Não é necessariamente elemento que faça parte do conjunto, para bem representá-lo, embora pertença obrigatoriamente ao intervalo entre a maior e a menor observação.

3) Não pode ser calculada para distribuições com limites indeterminados (indefinidos).

EXEMPLO: Idades de indivíduos submetidos a um teste

Idades	f_i
menos de 33	1
33 |— 35	21
35 |— 37	52
37 |— 39	186
39 |— 41	38
mais de 41	2
Total	300

Desprezando as classes com limites indeterminados, podemos calcular μ.

1.1.2. MÉDIA GEOMÉTRICA (OU MÉDIA LOGARITMICA)

Notação: M_g (popul.) $\bar{x}_g$ (am.)

$$M_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

Por logaritmo:

$$\log M_g = \frac{1}{N} \sum_{i=1}^{N} \log x_i \Rightarrow M_g = \text{antilog} \frac{\sum_{i=1}^{N} \log x_i}{N}$$

EXEMPLO:

x_i	$\log x_i$
120	2,07918
115	2,06070
120	2,07918
130	2,11394
120	2,07918
115	2,06070
Σ	12,47288

$$\log M_g = \frac{1}{6}\ 12,47288 \cong 2,07881 \Rightarrow M_g = \text{antilog}\ 2,07881 \cong 119,9$$

A média geométrica é o termo central de uma Progressão Geométrica. EXEMPLO: :: 3: 6: 12: 24: 48

MÉDIA GEOMÉTRICA PONDERADA : quando é calculada para um conjunto de dados dispostos numa distribuição de frequência.

$$M_g = \sqrt[N]{\prod_{i=1}^{m} x_i^{f_i}}$$

EXEMPLO:

$$M_g = \sqrt[6]{115^2 \times 120^2 \times 130^1} = 119,9$$

PROPRIEDADES:

P1: O produto dos quocientes de cada valor de um conjunto de números pela média geométrica do conjunto é igual a 1.

$$\frac{x_1}{M_g} \times \frac{x_2}{M_g} \times \dots \times \frac{x_N}{M_g} = 1$$

Demonstração:

$$\frac{x_1}{M_g} \times \frac{x_2}{M_g} \times \dots \times \frac{x_N}{M_g} = \frac{x_1 \quad x_2 \quad \dots \quad x_N}{(M_g)^N}$$

Extraindo a raiz n-ésima:

$$\frac{\sqrt[N]{x_1 \quad x_2 \quad \dots \quad x_N}}{\sqrt[N]{(M_g)^N}} = \frac{M_g}{M_g} = 1$$

EXEMPLO: Os valores da variável X são 10, 60 e 360 então:

$$M_g = \sqrt[3]{10 \times 60 \times 360} = 60 \quad \text{e} \quad \frac{10}{60} \times \frac{60}{60} \times \frac{360}{60} = \frac{216000}{216000} = 1$$

P2: Séries que apresentam o mesmo número de elementos com a mesma soma total tem a mesma média aritmética, enquanto séries que apresentam o mesmo número de elementos com o mesmo produto tem a mesma média geométrica.

EXEMPLO: Comprove esta propriedade tomando X={8;12,5} e Y={2;20;12}.

P3: A média geométrica é menor ou igual à média aritmética.

$$M_g \leq \mu$$

OBS: Esta propriedade sempre se verifica quando os valores da série forem positivos e nem todos iguais. Se entre eles houver um ou mais zeros, a média geométrica será nula.

A igualdade só ocorrerá quando todos os valores da série forem iguais.

Demonstração:Consideremos uma série com apenas dois valores:

$$(x_1 - x_2)^2 \geq 0 \qquad x_1, x_2 > 0$$

Desenvolvendo o primeiro membro:

$$x_1^2 + x_2^2 - 2\,x_1\,x_2 \geq 0$$

Somando e subtraindo a quantidade $2\,x_1x_2$ ao primeiro membro da desigualdade, teremos:

$$x_1^2 + x_2^2 + 2\,x_1\,x_2 - 2\,x_1\,x_2 - 2\,x_1\,x_2 \geq 0$$

ou $\quad (x_1 + x_2)^2 - 4\,x_1\,x_2 \geq 0 \quad$ e $\quad (x_1 + x_2)^2 \geq 4\,x_1\,x_2$

Daí tiramos

$$\frac{(x_1 + x_2)^2}{4} \geq x_1\,x_2 \quad \text{ou} \quad \left[\frac{x_1 + x_2}{2}\right]^2 \geq x_1\,x_2$$

Extraindo a raiz quadrada:

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 + x_2} \quad \text{,isto é:} \quad \mu \geq M_g \,.$$

P4: Quanto maior a diferença entre os valores originais maior será a diferença entre as médias aritmética e geométrica.

EXEMPLO: Consideremos os seguintes conjuntos de números:

X = { 2, 2}	Y = { 14, 16}	Z = { 8,12}	W = {2, 50}
$\mu_x = 2$	$\mu_y = 15$	$\mu_z = 10$	$\mu_w = 26$
$M_{g_x} = 2$	$M_{g_y} = 14{,}97$	$M_{g_z} = 9{,}80$	$M_{g_w} = 10$

APLICAÇÕES:

I- Média de relações.

Consideremos os dados a seguir:

Empresa	Capital líquido	Dívida	Relação capital/dívida	Relação dívida/capital
A	2 500	1 000	2,5	0,4
B	1 000	2 000	0,5	2,0
	3 500	3 000		

Calculando a média aritmética das duas relações, teremos:

Relação capital/dívida : $\mu_1 = \dfrac{2{,}5 + 0{,}5}{2} = 1{,}5$

Relação dívida/capital: $\mu_2 = \dfrac{0{,}4 + 2{,}0}{2} = 1{,}2$

Como essas relações são o inverso uma da outra, é um contra-senso que o produto de suas médias não seja igual a 1, pois: $(1{,}5 \times 1{,}2) > 1$. Entretanto esse aparente absurdo ocorre do fato de não havermos atribuído às relações seus pesos corretos antes de calcularmos a média. Teríamos o seguinte resultado, se os pesos houvessem sido considerados:

$$\frac{(2{,}5 \times 1000) + (0{,}5 \times 2000)}{1000 + 2000} = 1{,}1667$$

O numerador dessa expressão corresponde ao capital líquido total (3500), enquanto que o denominador é igual à dívida total (3000). O resultado 1,1667 deve ser interpretado como a relação média entre capital e dívida, entre as duas empresas consideradas.

Analogamente, determinamos a média das relações da dívida versus capital líquido:

$$\frac{(0{,}4 \times 2500) + 2{,}0 \times 1000)}{2500 + 1000} = 0{,}8571$$

ou, com o emprego dos totais: $\frac{3000}{3500} = 0{,}8571$

Multiplicando as duas médias, obtemos:

$1{,}1667 \times 0{,}8572 = 1$

Conforme se pode verificar, os pesos atribuídos às relações não foram iguais. Caso se deseje que isso ocorra, pode-se recorrer à média geométrica. Assim:

Relação capital/dívida: $M_{g_1} = \sqrt{2{,}5 \times 0{,}5} = 1{,}1180$

Relação dívida/capital: $M_{g_2} = \sqrt{0{,}4 \times 2{,}0} = 0{,}8944$

O produto dessas médias é 1.

A escolha entre as médias dependerá, nesse e em outros casos, do fim que se persegue. Se, para uma determinada empresa, se deseja estabelecer uma relação do tipo capital/dívida que seja independente da dívida ou do capital das diferentes empresas envolvidas, é recomendável o uso da média geométrica. Se o que se deseja é a relação capital líquido/dívida de um certo número de empresas, após a consolidação, a cifra correta será obtida através da média aritmética, ou achando a relação dos totais.

II- Médias em distribuições assimétricas.

Uma distribuição de frequência pode

ser inclinada à direita (assimétrica). Se usarmos os logaritmos dos valores da variável, com um intervalo de classe constante para os logaritmos, a curva se transformará em simétrica. Neste caso, a média geométrica revela-se mais apropriada que a aritmética.

III- Média de taxas de variação.

Em certas ocasiões, a média geométrica é utilizada para determinar taxas médias. Suponhamos que um indivíduo tenha investido um capital de Cr$ 500 em 1983. Após um ano de aplicação, essa importância passou para Cr$ 650. Reaplicando essa última quantia, ao final de mais um ano seu montante passou para Cr$ 910. A taxa média de aumento de capital será obtida mediante o cálculo de uma média geométrica.

Calculemos, inicialmente, as taxas de aumento de capital, período a período:

Período	Taxa
1983-1984	$\frac{650}{500} = 1,3$
1984-1985	$\frac{910}{650} = 1,4$

Taxa média: $\sqrt{1,3 \times 1,4} = 1,3491$

A taxa média de aumento do capital investido no período de dois anos foi de 34,91%.

IV- Estimativa do crescimento demográfico.

V- Cálculo do índice do custo de vida.

Tem ampla aplicação na Estatística Econômica e Demografia em virtude de possibilitar cálculo de médias de razões ou coeficientes de aumento ou diminuição.

1.1.3. MÉDIA HARMÔNICA

Notação: M_h (popul.) $\bar{x}_h$ (am.)

Definição: É o recíproco da média aritmética dos recíprocos dos termos.

$$M_h = \frac{1}{\dfrac{\sum_{i=1}^{N} \dfrac{1}{x_i}}{N}} = \frac{N}{\sum_{i=1}^{N} \dfrac{1}{x_i}}$$

A origem da denominação de média harmônica é musical. Esse é o termo central da sucessão 6, 4, 3 onde 6:4:3 é a sequência segundo a qual deve estar o comprimento da corda musical para se obter uma nota, a quinta e a oitava. De fato:

$$\frac{1}{\frac{1}{3}\left(\frac{1}{6} + \frac{1}{4} + \frac{1}{3}\right)} = \frac{3}{\frac{9}{12}} = 4$$

sendo que $\mu = \frac{13}{3} = 4{,}33$ e $M_g = \sqrt[3]{72} = 4{,}16$

PROPRIEDADE: A média harmônica é menor ou igual à média geométrica para valores da variável diferentes de zero.

$$M_h \leq M_g \leq \mu$$

OBS: Quando os valores não forem muito diferentes, verifica-se aproximadamente a seguinte relação:

$$M_g = \frac{\mu + M_h}{2}$$

EXEMPLO:

Valores	μ	M_g	M_h
X={2,4,6,8,10}	6	5,21	4,38
Y={10,10,10}	10	10	10
Z={10,1;10,1;10,2; 10,4;10,5}	10,26	10,2587	10,2574

MÉDIA HARMÔNICA PONDERADA:

$$M_h = \frac{\sum_{i=1}^{m} f_i}{\sum_{i=1}^{m} \frac{1}{x_i} f_i} = \frac{N}{\sum_{i=1}^{m} \frac{f_i}{x_i}}$$

APLICAÇÕES:

I- Séries contínuas de números recíprocos, tais como: cálculo do custo médio de artigos adquiridos com uma determinada quantia e cálculo das médias de velocidades referidas a uma dada distância.

EXEMPLOS:

(a) Uma pessoa adquiriu quatro camisas ao preço unitário de Cr$ 30 000 e duas camisas a Cr$ 50 000 cada uma.
Qual o preço médio paga por camisa ?

(1)

$$M_h = \frac{1}{\frac{120000(1/30000) + 100000(1/50000)}{120000 + 100000}} = 36\ 670$$

pois a pessoa gastou 120000 em camisas de 30000 e 100000 em camisas de 50000. Nesse caso, os valores da variável são os preços por camisa e seus pesos as quantidades gastas em camisas.

(2)

$$\mu = \frac{(4 \times 30000) + (2 \times 50000)}{4 + 2} = 36\ 670$$

Os valores da variável são os preços por camisa e os pesos são o número de camisas compradas.

O princípio dos dois métodos é simples. Estamos calculando média de taxas, ou seja, cruzeiros por camisa. O numerador de ambas as expressões representa a quantia total gasta com camisas (220000), enquanto o denominador é igual ao número de camisas adquiridas. Quando são empregados os pesos do denominador, aplica-se a média aritmética (pesos 4 e 2); quando se empregam os pesos do numerador, usa-se a média harmônica (120000 e 100000).

(b) Um automóvel parte do Rio de Janeiro com destino na São Paulo desenvolvendo uma velocidade média de 70 km/h, voltando no dia seguinte ao ponto de partida com uma velocidade média de 90 km/h. Qual será a velocidade média considerando a viagem completa (ida e volta) ?

A média não é 80 km/h e sim:

$$M_h = \frac{2}{\frac{1}{70} + \frac{1}{90}} = 78,75 \text{ km/h}$$

pois, suponhamos que entre Rio e São Paulo a distância seja de 450 km (o que não influirá no raciocínio). Sabemos que a média é a relação entre a distância total percorrida e o tempo total gasto para percorrê-la. O carro andou 2 x 450km = 900 km. Se o carro na ida desenvolveu 70 km/h, ele levou (450 km)/(70 km/h) para chegar a S.P. e na volta (450 km)/(90 km/h). Assim:

$$\text{velocidade média} = \frac{\text{distância total}}{\text{tempo total}} = \frac{2 \times 450}{\frac{450}{70} + \frac{450}{90}} = 78,75\text{km/h}$$

II- Grande valia quando os dados apresentam uma relação de reciprocidade.

EXEMPLO: Suponhamos que um cidadão tenha contratado o serviço de dois operários para construir uma parede em sua residência e que essa tarefa implique na colocação de 900 tijolos. Sabendo que o operário A coloca 400 tijolos em 8 h de trabalho e B coloca 500 em 6 h, quanto tempo será dispendido pelos dois operários para assentarem os 900 tijolos ?

A produtividade dos operários é diferente, razão porque, se usarmos μ estaremos nivelando as capacidades individuais.

Temos: 8 horas $\Rightarrow$ 8 x 60 = 480 min

A $\Rightarrow$ 480 : 400 = 1,2 min por tijolo

6 h $\Rightarrow$ 6 x 60 = 360 min

B $\Rightarrow$ 360 : 500 = 0,72 min por tijolo

$$M_h = \frac{2}{\frac{1}{1,2} + \frac{1}{0,72}} = 0,9 \text{ min por tijolo}$$

$900 \times 0{,}9 = 810$ min (ao todo)

Cada operário gastará: $810 : 2 = 405$ min $= 6$ h 45 min ,

sendo que A colocará $405 : 1{,}2 = 337{,}5$ tijolos e

B colocará $405 : 0{,}72 = 562{,}5$ tijolos.

VANTAGEM: Pouco influenciada por valores extremos.

1.1.4. MÉDIA DE POTÊNCIA E MÉDIA QUADRÁTICA

Notação: M_p (pop.) $\bar{X}_p$ (am.)

Fórmula geral:

$$M_p = \sqrt[p]{\frac{\sum^{N} x_i^p}{N}}$$

onde x_i é positivo

$i = 1, \ldots, N$

$p \in Z^*$

$$p = 1 \Rightarrow \mu = \sqrt[1]{\frac{\sum x_i}{N}} = \frac{\sum x_i}{N}$$

$$p = -1 \Rightarrow M_h = \sqrt[-1]{\frac{\sum x_i^{-1}}{N}} = \frac{N}{\sum \frac{1}{x_i}}$$

$$p = 2 \Rightarrow M_q = \sqrt[2]{\frac{\sum x_i^2}{N}} \Rightarrow \text{média quadrática}$$

para dados agrupados :
(média quadrática ponderada) $M_q = \sqrt[2]{\frac{\sum f_i x_i^2}{N}}$

$$p = 3 \Rightarrow M_c = \sqrt[3]{\frac{\sum x_i^3}{N}} \Rightarrow \text{média cúbica}$$

$$p = 4 \Rightarrow M_{bq} = \sqrt[4]{\frac{\sum x_i^4}{N}} \Rightarrow \text{média biquadrática}$$

PROPRIEDADES DA MÉDIA QUADRÁTICA:

P1: A média quadrática de uma constante é igual à constante.

P2: Multiplicando ou dividindo todos os valores de um conjunto de números por uma constante , a média quadrática fica multiplicada ou dividida por essa constante.

P3: Sempre que os valores de X forem positivos é válida a relação:

$$X_{min} < M_h \leq M_g \leq \mu \leq M_q \leq X_{max}$$

A igualdade só se verifica quando os valores da variável forem iguais.

APLICAÇÕES: A média quadrática tem aplicação na Física e é largamente utilizada em Estatística, principalmente quando se pretende calcular a média de desvios $(x - \mu)$, em vez de a média dos valores originais. Neste caso, a média quadrática é denominada desvio padrão, que é uma importante medida de dispersão a ser estudada mais adiante.

1.2. SEPARATRIZES.

As separatrizes são valores de posição que dividem o conjunto em partes iguais.

Para o cálculo das separatrizes é necessário, primeiramente, dispor os dados em ordem crescente ou decrescente.

1.2.1. MEDIANA.

Notação: M_e

É o valor central.

EXEMPLOS : a) QI de 5 indivíduos: 93 94 97 102 102

A Mediana é 97 pois existem à esquerda do mesmo tantos valores quanto à direita. A mediana ocupa, neste caso, o 3º lugar, sendo sua posição indicada por:

$$P = \frac{N + 1}{2}, \text{ onde N é ímpar.}$$

b) QI de 12 indivíduos : 93 94 95 97 98 98 100 101 101 104 105 108

Quando N é par, a mediana é dada por:

$$M_e = \frac{X_P + X_{P+1}}{2} \quad \text{onde} \quad P = \frac{N}{2}$$

isto é,a mediana é a média aritmética dos valores centrais.

Então:

$$P = \frac{12}{2} = 6 \quad \Rightarrow \quad M_e = \frac{X_6 + X_7}{2} = \frac{98 + 100}{2} = 99$$

Em se tratando de dados agrupados, calcula-se as frequências acumuladas a fim de localizar a posição da mediana obtida por N/2, independente de N ser par ou ímpar, pois quando trabalhamos com um grande número de dados, a inclusão de 1 pouco modifica o resultado.

EXEMPLO:

Tabela 2: Estatura de 50 Alunos
Escola Êxitus - P.A. - 1992

cm	f_i	F_i
160 |— 164	6	6
164 |— 168	8	14
168 |— 172	9	23
172 |— 176	17	40
176 |— 180	6	46
180 |— 184	2	48
184 |— 188	2	50
Total	50	

A posição N/2 = 50/2 = 25. Verifica-se então que M_e localiza-se na $4^{\underline{a}}$ classe e será igual ao limite inferior 172 acrescida de uma certa quantidade Δ . Assim: $M_e = 172 + \Delta$, pois $172 < M_e < 176$. Ora, até a $3^{\underline{a}}$ classe, temos 23 ocorrências e até a $4^{\underline{a}}$, 40; como a mediana está no $25^{\underline{o}}$ lugar podemos considerar 25 - 23 = 2 casos na $4^{\underline{a}}$ classe. Admite-se que os dados se distribuem uniformemente no intervalo de classe e como na $4^{\underline{a}}$ classe temos 17 casos, devemos calcular o intervalo correspondente a 2 casos:

$$\begin{array}{ccc} 17 & \_\_ & 4 \\ 2 & \_\_ & \Delta \end{array} \quad \Rightarrow \quad \Delta = \frac{2 \times 4}{17} = 0{,}47$$

$$M_e = 172 + 0{,}47 = 172{,}47 \text{ cm}$$

Podemos estabelecer a fórmula através do raciocínio:

1º) M_e situa-se entre os limites inferior e superior da classe que a contém;

2º) N/2 situa-se entre as F_i anterior à classe da mediana e à classe que contém a mediana;

Assim:

$$l_i < M_e < l_s$$
$$F_{ant} < N/2 < F_i$$

$$\frac{M_e - l_i}{l_s - l_i} = \frac{N/2 - F_{ant}}{F_i - F_{ant}} \Rightarrow \frac{M_e - l_i}{h} = \frac{N/2 - F_{ant}}{f_i}$$

$$M_e - l_i = h\left[\frac{N/2 - F_{ant}}{f_i}\right] \Rightarrow M_e = l_i + h\left[\frac{N/2 - F_{ant}}{f_i}\right]$$

onde l_i = limite inferior da classe mediana

f_i = frequência absoluta simples da classe mediana

DETERMINAÇÃO GRÁFICA DA MEDIANA:

A mediana pode ser facilmente encontrada a partir da ogiva. A mediana é a abcissa do ponto P da ogiva cuja ordenada é o elemento mediano, no caso de frequências absolutas e 50%, no caso de frequências relativas.

VANTAGENS E DESVANTAGENS:

1) Não depende de todos os valores da série, podendo até mesmo não se alterar com a modificação de alguns deles.

2) Não é influenciada pelos valores extremos.

3) Pode ser calculada para distribuições com limites indeterminados, na maioria dos casos.

4) É elemento que faz parte do conjunto para N ímpar e dados não agrupados.

APLICAÇÕES: É indicada principalmente para distribuições assimétricas.

Tem ampla aplicação em questões educacionais.

Utilizada para variáveis medidas em escala ordinal.

1.2.2. QUARTIS

Os quartis dividem os conjuntos ordenados em 4 partes iguais: 25% dos valores são inferiores ao 1º Quartil (Q1); 50% ao 2º Quartil ($Q2 = M_e$); 75% inferiores ao 3º Quartil (Q3).

Posição do Q1 : $\dfrac{N + 1}{4}$

Posição do Q2 : $\dfrac{2(N + 1)}{4} = \dfrac{N + 1}{2}$

Posição do Q3 : $\dfrac{3(N + 1)}{4}$

Exemplo: 1, 3, 5, 6, 8, 9, 12, 15, 17, 19, 21, 25, 28, 29, 31, 35, 36, 40, 41

Posição do Q1 = $\dfrac{19 + 1}{4}$ = 5º lugar $\Rightarrow$ Q1 = 8

Posição do Q3 = $\dfrac{3(19+1)}{4}$ = 15º lugar $\Rightarrow$ Q3 = 31

Em se tratando de um número par de valores, a mediana e os quartis ficarão situados entre valores, podendo-se então usar a fórmula:

$M_e = l_i + (L_i - l_i)\alpha$ onde α é a parte fracionária resultante da divisão (N+1)/2

$Q1 = l_i + (L_i - l_i)\alpha$ onde α é a parte fracionária resultante da divisão (N+1)/4.

EXEMPLO: 3 , 7 , 8 , 10 , 12 , 15

$P_{M_e} = (N+1)/2 = 3{,}5^{\underline{o}}$ $\Rightarrow$ $M_e = (8+10)/2 = 9$ ou

$M_e = 8 + (10-8).0{,}5 = 9$

$P_{Q1} = (N+1)/4 = (6+1)/4 = 1{,}75^{\underline{o}}$ $\Rightarrow$ $Q1 = 3 + (7-3).0{,}75 = 6$

Os quartis para dados agrupados são obtidos pela fórmula da mediana, bastando substituir N/2 por N/4 ou 3N/4.

pela fórmula da mediana, bastando substituir N/2 por N/4 ou 3N/4.
EXEMPLO: Considere a Tabela 2.

$$Q_1 = l_i + h\left[\frac{N/4 - F_{ant}}{f_i}\right] = 164 + 4\left[\frac{12,5 - 6}{8}\right] = 167,25$$

$$Q_3 = l_i + h\left[\frac{3N/4 - F_{ant}}{f_i}\right] = 172 + 4\left[\frac{37,5 - 23}{17}\right] = 175,41$$

1.2.3. DECIS.

Dividem os conjuntos ordenados de 10 em 10.

Para dados agrupados são obtidos também pela fórmula da mediana, substituindo N/2 por N/10, 2N/10, etc.

EXEMPLO: Tomando novamente a Tabela 2:

$$D_1 = l_i + h\left[\frac{N/10 - F_{ant}}{f_i}\right] = 160 + 4\left[\frac{5 - 0}{6}\right] = 163,33$$

$$D_2 = l_i + h\left[\frac{2N/10 - F_{ant}}{f_i}\right] = 164 + 4\left[\frac{10 - 6}{8}\right] = 166$$

que são os $1^{\underline{o}}$ e $2^{\underline{o}}$ decis, respectivamente.

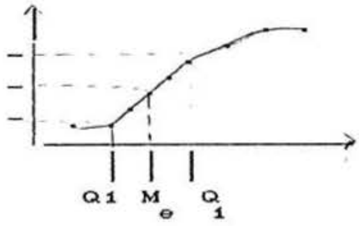
1.2.4. PERCENTIS.

Dividem os conjuntos ordenados de 100 em 100.

Também se utiliza a fórmula da mediana para dados agrupados, com a respectiva adaptação: substitui-se N/2 por N/100, 2N/100, 3N/100, etc.

DETERMINACÃO GRÁFICA DAS SEPARATRIZES:

As separatrizes podem ser obtidas a partir da ogiva analogamente à mediana.



1.3 DOMINANTES.

São os valores calculados em um conjunto em função das respectivas frequências.

1.3.1. MODA.

Notação: M_o

O dado que aparece o maior número de vezes, tendo, portanto, a maior frequência, é a Moda, também chamada de Norma, Valor Dominante ou Valor Típico.

Um conjunto com duas modas é chamado bimodal e com mais de duas modas , polimodal.

Um conjunto sem moda é chamado antimodal. Por exemplo, séries em forma de U.

EXEMPLO : 3, 5, 8, 9, 11 não tem moda

3, 5, 8, 8, 9, 11 $M_o = 8$

3, 5, 8, 8, 9, 9, 11 $M_o = 8$ e 9

Para dados agrupados, podemos calcular:

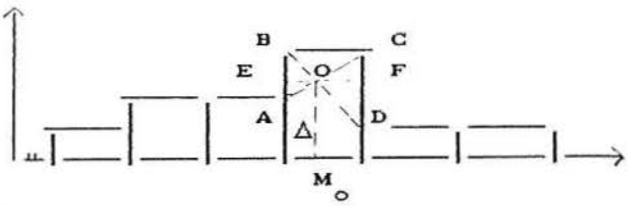
1º) Moda bruta.

É o ponto médio da classe de maior frequência.

EXEMPLO: A moda bruta da Tabela 2 é 174 cm.

2º) Moda pelo processo de Czuber

A fórmula de Czuber é calculada em função da classe modal e das frequências a ela adjacentes.



$M_o = l_i + \Delta$

Os triângulos AOB e COD são semelhantes, logo:

$$\frac{AB}{CD} = \frac{OE}{OF} \quad , \text{ ora:}$$

$AB = f_i - f_{ant}$; $CD = f_i - f_{post}$; $OE = \Delta$; $OF = h - \Delta$.

Temos então:

$$\frac{f_i - f_{ant}}{f_i - f_{post}} = \frac{\Delta}{h - \Delta}$$

$$(f_i - f_{ant})\ (h - \Delta) = \Delta\ (f_i - f_{post})$$

$$f_i\, h - f_i\, \Delta - h\, f_{ant} + f_{ant}\, \Delta = \Delta\, f_i - \Delta\, f_{post}$$

$$h\ (f_i - f_{ant}) = \Delta\ (2\, f_i - f_{post} - f_{ant})$$

$$\Delta = h\ \frac{f_i - f_{ant}}{2\, f_i - (f_{post} + f_{ant})} \quad \rightarrow$$

$$M_o = l_i + h\ \frac{f_i - f_{ant}}{2\, f_i - (f_{post} + f_{ant})}$$

onde l_i = limite inferior da classe modal;

h = amplitude do intervalo;

f_i = frequência da classe modal;

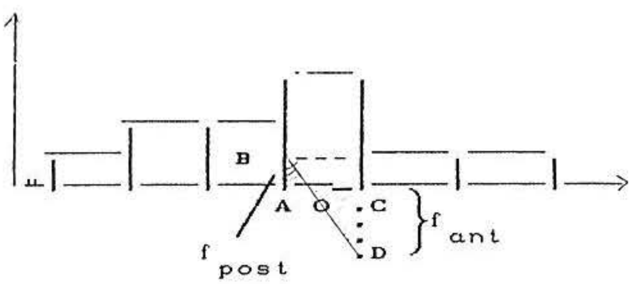
f_{ant} = frequência da classe antecedente à classe modal;

f_{post} = frequência da classe posterior à classe modal.

EXEMPLO: Relativamente aos dados da TAbela 2:

$$M_o = 172 + 4\left[\frac{17 - 9}{34 - (9+6)}\right] = 173{,}68 \text{ cm.}$$

3º) Moda pelo processo de King



Os triângulos AOB e COD são semelhantes, logo:

$$\frac{AB}{AO} = \frac{CD}{CO}$$

$AB = f_{post}$ $\qquad AO = \Delta$ $\qquad CD = f_{ant}$

$$\frac{f_{post}}{\Delta} = \frac{f_{ant}}{h - \Delta} \Rightarrow f_{post} (h - \Delta) = f_{ant} \Delta$$

$$h f_{post} = \Delta f_{ant} + \Delta f_{post} \Rightarrow$$

$$h f_{post} = \Delta (f_{ant} + f_{post}) \Rightarrow$$

$$\Delta = h \frac{f_{post}}{f_{ant} + f_{post}}$$

$$M_o = l_i + h \frac{f_{post}}{f_{ant} + f_{post}}$$

EXEMPLO: Na Tabela 2: $M_o = 172 + 4 \dfrac{6}{9 + 6} = 173{,}6$

4º) Moda pelo processo de Pearson

Karl Pearson estabeleceu a fórmula:

$$\mu - M_o = 3 (\mu - Me) \Rightarrow M_o = 3 M_e - 2 \mu$$

que fornece bons resultados se a curva não for muito assimétrica.

APLICAÇÃO: Uma fábrica ocupa mensalmente 72 operários de janeiro a outubro, mas nos últimos meses, tendo em vista o mercado favorável, contrata mais operários, empregando 102 em novembro e 150 em dezembro. Neste caso,

$\mu = \dfrac{10 \cdot 72 + 150 + 102}{12} = 81$. Ora, em nenhum mes a fábrica empregou 81 operários; o valor 72 é mais representativo, pois na verdade, a fábrica trabalhou mensalmente com 72 operários e só ocasionalmente precisou aumentar seu número, o suficiente para influenciar a média. Nesse caso, a moda é muito mais representativa.

VANTAGENS E DESVANTAGENS:

1) Não depende de todos os valores da série, nem de sua ordenação (rol), podendo até não se modificar se alterar algum valor.

2) Não é influenciada por valores extremos da série.

3) Sempre é representada por um elemento do conjunto de dados, exceto o caso de distribuição de frequência por classe, quando trabalhamos com subconjuntos (dados agrupados) e não com cada elemento isoladamente.

4) Pode ser calculada para distribuições com limites indeterminados, na maioria dos casos.

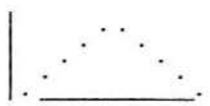
5) Não é única e pode não existir para certos conjuntos.

RELACÃO ENTRE MÉDIA, MEDIANA E MODA.

Para dados agrupados em distribuição de frequência, as diferenças entre μ , M_e e M_o são indicadores da forma da curva em termos de assimetria.

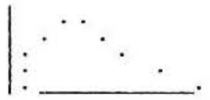
Para uma distribuição unimodal simétrica:

$\mu = M_e = M_o$



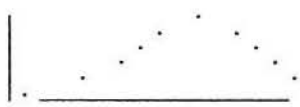
Para uma curva positivamente assimétrica:

$M_o < M_e < \mu$



Para uma curva negativamente assimétrica:

$\mu < M_e < M_o$



1.3.2. ANTIMODA.

O valor que ocorre com a menor frequência é a antimoda.

1.3.3. VALOR PREVALENTE.

Efetuando-se o produto dos dados pelas respectivas frequências, chamamos de valor prevalente ou valor dominante ao maior desses produtos.

1.4. EXERCÍCIOS.

1. Aplicou-se um teste de raciocínio abstrato a dois grupos de 35 e 38 alunos, respectivamente, sendo a média aritmética do primeiro grupo 11,8 e a do segundo grupo 12,7. Qual é a média aritmética dos dois grupos considerados conjuntamente?

2. Uma partida de carne destinada à merenda escolar de certa região administrativa foi pesada 10 vezes, obtendo-se os resultados,em toneladas:1325,29 1325,28 1325,30 1325,30 1325,28 1325,29 1325,27 1325,29 1325,29 1325,31. Qual é o peso médio que mais se aproxima do real? Justifique

3. Tendo-se aplicado a um grupo de alunos dois testes de Aptidão Verbal, verificou-se média 5 no 1º teste e 6 no 2º. Será possível, sem se conhecer o número de alunos, calcular a média dos dois testes ? E, se considerarmos que o grupo é formado por 100 alunos ?

4. Calcular o tempo médio, sabendo que vários capitais empregados à mesma taxa rendem o mesmo juro durante os tempos:

Tempo (meses)	No.de capitais
1 |— 3	10
3 |— 5	22
5 |— 7	30
7 |— 9	28
9 |— 11	10

5. Dois candidatos submeteram-se a um concurso obtendo os seguinte resultados:

Matéria	Peso	Candidato A	Candidato B
Contabilidade	2	7,0	6,5
Matemática	2	8,0	6,0
Portugues	1	5,0	7,0
Datilografia	1	6,0	7,0
Legislação	2	5,0	6,0

Classificar os candidatos caso o critério seja:

a) μ b) M_h c) M_g d) M_q .

6. O professor de Educação Física procedeu a um levantamento a respeito das estaturas de seus alunos, verificando que, das 22 medições efetuadas, as alturas situaram-se entre 148 cm e 182 cm, acusando média aritmética igual a 164 cm. Posteriormente, ele constatou que a fita métrica utilizada apresentava 3 cm a menos. Em vista da retificação que deverá ser feita, pergunta-se:

a) será necessário medir novamente todos os alunos ?

b) quais os novos limites superior e inferior das alturas ?

c) qual a nova média ?

7. A soma dos pesos dos componentes de uma equipe de vôlei é 648 kg. A relação entre o número de componentes e a média aritmética é 1/8. Determinar o número de componentes do grupo e a média dos pesos.

8. Foi calculada a média aritmética dos QI de 8 indivíduos, encontrando-se 78 pontos. Sabendo que os resultados foram 85 68 92 64 80 76 86 73, verifique se a média está correta, sem calculá-la pela fórmula.

9. Em certa fábrica de laticínios, a máquina de empacoter leite está apresentado defeito, para mais, na ordem de 5 % em cada litro. Em média são vendidos cerca de 54oo litros de leite diàriamente. Determinar o prejuízo, em litros diários, que a empresa sofre e a média diária que a mesma deveria vender.

10. Os escores a seguir referem-se aos resultados de sujeitos de um Grupo Experimental (GE) e de um Grupo de Controle (GC) num teste de raciocínio lógico. O GE foi submetido a um nível de grande ruído e o GC a um nível de ruído mínimo.

GE (x_i)	f_i	GC (y_i)	f_i
5	1	9	2
6	1	10	2
7	1	11	3
8	2	12	5
9	3	13	4
10	4	14	3
11	2	15	1
12	3		
13	2		
14	1		

Calcule a mediana e os quartis de cada grupo e interprete.

11. Uma corporação foi dividida, segundo a altura de seus membros, em quatro grupos de igual número de elementos, para a realização de exercícios físicos. Quais as alturas limites para cada grupo ?

Altura (cm)	Nº de indivíduos
150 |— 155	4
155 |— 160	16
160 |— 165	20
165 |— 170	33
170 |— 175	27
175 |— 180	15
180 |— 185	5

12. Para uma amostra de 15 clientes de um pequeno mercado foram observados os seguintes montantes de vendas, em ordem crescente: 10 10 25 25 25 35 40 53 90 125 135 245 271 309 410 (u.m.).

a) Determinar:a.1) média aritmética a.2) mediana a.3)moda

b) Como voce descreveria esses dados do ponto de vista de assimetria ?

c) Se lhe pedissem uma descrição dos dados que envolvesse a informação da quantidade "típica" de compra por cliente, qual a medida de tendência central que voce usaria ?

13. Comentar a forma da distribuição das taxas de aluguel de apartamentos, a partir dos cálculos e interpretações de:

a) média aritmética; b) moda;c) mediana; d) 1º e 3º quartis; e) 1º e 2º decis.

Aluguéis mensais de apartamentos
Bairro Floresta - P.A. - 1990

mil cruzeiros	No.de apartamentos
150 |— 180	3
180 |— 210	8
210 |— 240	10
240 |— 270	13
270 |— 300	33
300 |— 330	40
330 |— 360	35
360 |— 390	30
390 |— 420	16
420 |— 450	12

Fonte: Imobiliária MOREBEM

14. Aplicado um teste de momória a dois grupos A e B de 128 indivíduos cada um, de mesma faixa etária, calculou-se:

A: $\mu = 54{,}8$ $Q_1 = 44{,}4$ $Q_2 = 54{,}6$ $Q_3 = 63{,}6$

B: $\mu = 53{,}2$ $Q_1 = 42{,}6$ $Q_2 = 54{,}6$ $Q_3 = 62{,}2$

a) Faça um esquema grágico dos grupos.

b) Pela comparação das medidas, que conclusões voce tiraria?

15. Idade dos alunos do Turno da Noite da Escola Til

Idade	f_i	F_i
18 |— 20	120	120
20 |— 22	150	270
22 |— 24	180	450
24 |— 26	200	650
26 |— 28	190	840
28 |— 30	160	1000

Fonte: Secretaria da Escola

Calcule e interprete: a) média aritmética; b)moda de Czuber c) antimoda; d) valor prevalente; e) mediana; f) 2º quartil

2. CARACTERÍSTICAS DE VARIABILIDADE

Uma descrição completa dos dados exige que, além da apresentação através de tabelas e gráficos e do cálculo de medidas de tendência central, tenhamos informações quanto à sua variabilidade.

Suponhamos, por exemplo, que se deseja comparar a performance de dois empregados, com base na produção diária de determinada peça:

Empregado A: 70, 71, 69, 70, 70 $\Rightarrow \mu_A = 70$

Empregado B: 60, 80, 70, 62, 83 $\Rightarrow \mu_B = 71$

Baseados nestes únicos resultados, diríamos que a performance de B é melhor do que a de A, já que B produz, em média, um maior número de peças diàriamente. No entanto, se formos um pouco mais cuidadosos, perceberemos que a produção de A varia apenas de 69 a 71 peças, ao passo que a de B varia de 60 a 83 peças, o que indica que a performance de A é bem mais uniforme do que a de B. Ocorre, por outro lado, que um alto grau de uniformidade ou dispersão costuma ser considerado como uma qualidade desejável em um processo produtivo. Qualquer produção em série seria antieconômica se houvesse muita variabilidade nos materiais ou peças fabricadas.

2.1. MEDIDAS DE DISPERSÃO ABSOLUTA

Consideremos os seguintes conjuntos de dados que ilustrarão o desenvolvimento de nosso conteúdo:

	Conjunto 1	Conjunto 2	Conjunto 3	Conjunto 4	Conjunto 5
	5	4	2	0	4
	5	5	4	2	4
	5	5	5	5	4
	5	5	6	8	5
	5	6	8	10	6
					6
					6
μ	5	5	5	5	5
M_e	5	5	5	5	5
Mo	não tem	5	não tem	não tem	4 e 6

Como podemos observar, os 5 conjuntos não diferem entre si se considerarmos sòmente a média e a mediana. Acrescentando a informação da moda, podemos diferenciar alguns deles. As medidas de variabilidade ou dispersão possibilitam que façamos distinção entre os conjuntos quanto à sua homogeneidade, isto é, o grau de concentração em torno de uma medida de tendência central.

2.1.1. AMPLITUDE

Notação: H

É a diferença entre o maior e o menor valor observado.

$$H = X_{máx} - X_{min}$$

EXEMPLO: $H_1 = 0$ $H_2 = 2$ $H_3 = 6$ $H_4 = 10$ $H_5 = 2$

2.1.2. DESVIO MÉDIO

Notação: DM

É a média aritmética dos valores absolutos dos desvios tomados em relação à média aritmética ou em relação à mediana.

2.1.2.1. DESVIO MÉDIO EM RELAÇÃO A μ

$$DM = \frac{\sum^{N} |d_i|}{N} = \frac{\sum^{N} |X_i - \mu|}{N} \quad \text{(dados não agrupados)}$$

$$DM = \frac{\sum^{m} f_i |d_i|}{N} = \frac{\sum^{m} f_i |X_i - \mu|}{N} \quad \text{(dados agrupados)}$$

EXEMPLO: Para os dados do Conjunto 3:

x_i	$x_i - \mu$	$\lvert x_i - \mu\rvert$
2	-3	3
4	-1	1
5	0	0
6	1	1
8	3	3
Σ 25	0	8

$\mu_3 = 5$

$$DM_3 = \frac{8}{5} = 1,6$$

Anàlogamente: $DM_1=0$ $DM_2=2/5=0,4$ $DM_4=16/5=3,2$ $DM_5=6/5=1,2$

EXEMPLO: Para dados em distribuição de frequência, temos:

Tabela 3: Idade das crianças matriculadas na Escola Tamborzinho - 92/1

Anos	f_i	x_i	$f_i x_i$	$x_i - \mu$	$\lvert x_i - \mu \rvert$	$f_i \lvert x_i - \mu \rvert$
0 |— 2	5	1	5	-3,6	3,6	18
2 |— 4	15	3	45	-1,6	1,6	24
4 |— 6	20	5	100	0,4	0,4	8
6 |— 8	5	7	35	2,4	2,4	12
8 |— 10	5	9	45	4,4	4,4	22
	50		230	/	/	84

$\mu = 230/50 = 4{,}6$ anos DM = 84/50=1,68 anos

2.1.2.2. DESVIO MÉDIO EM RELAÇÃO À M_e

$$DM_{M_e} = \frac{\sum^{N} |X_i - M_e|}{N} \quad \text{(dados não agrupados)}$$

$$DM_{M_e} = \frac{\sum^{m} f_i |X_i - M_e|}{N} \quad \text{(dados agrupados)}$$

EXEMPLO: Idade das crianças matriculadas na Escola Tamborzinho - 92/1

Anos	f_i	x_i	F_i	$x_i - M_e$	$\lvert x_i - M_e \rvert$	$f_i \lvert x_i - M_e \rvert$
0 |— 2	5	1	5	-3,5	3,5	17,5
2 |— 4	15	3	20	-1,5	1,5	22,5
4 |— 6	20	5	40	0,5	0,5	10,0
6 |— 8	5	7	45	2,5	2,5	12,5
8 |— 10	5	9	50	4.5	4,5	22,5
Σ	50	/	/	/	/	85,0

$$M_e = 4 + 2\,\frac{50/2 - 20}{20} = 4{,}5 \text{ anos} \Rightarrow DM_{M_e} = 85/50 = 1{,}7 \text{ anos}$$

2.1.3. SOMA DE QUADRADOS

Notação: SQ

É a soma dos quadrados dos desvios em torno de μ.

$$SQ = \sum^{N} d_i^2 = \sum^{N} (x_i - \mu)^2 \quad \text{(dados não agrupados)}$$

$$SQ = \sum^{m} f_i d_i^2 = \sum^{m} f_i (x_i - \mu)^2 \quad \text{(dados agrupados)}$$

EXEMPLO: Para os dados do Conjunto 3:

x_i	$x_i - \mu$	$(x_i - \mu)^2$
2	-3	9
4	-1	1
5	0	0
6	1	1
8	3	9
Σ 25	0	20

$\mu = 25/5 = 5$

$SQ_3 = 20$

Anàlogamente: $SQ_1 = 0$ $SQ_2 = 2$ $SQ_4 = 68$ $SQ_5 = 6$

EXEMPLO: Idade das crianças matriculadas na Escola Tamborzinho - 92/1

Anos	f_i	x_i	$f_i x_i$	$x_i - \mu$	$(x_i - \mu)^2$	$f_i (x_i - \mu)^2$
0 |— 2	5	1	5	-3,6	12,96	64,8
2 |— 4	15	3	45	-1,6	2,56	38,4
4 |— 6	20	5	100	0,4	0,16	3,2
6 |— 8	5	7	35	2,4	5,76	28,8
8 |— 10	5	9	45	4,4	19,36	96,8
Σ	50		230	/	/	232,0

$\mu = 4,6$ anos $SQ = 232,0$ anos^2

OBSERVAÇÃO 1: A soma de quadrados é aplicada exaustivamente na Análise de Variância.

OBERVAÇÃO 2: Como a SQ não leva em consideração o número de elementos do conjunto, definimos:

2.1.4. VARIÂNCIA ABSOLUTA

Notação : σ^2 (população) s^2 (amostra)

$$\sigma^2 = \frac{\sum^{N} d_i^2}{N} = \frac{\sum^{N} (x_i - \mu)^2}{N} \quad \text{(dados não agrupados)}$$

$$\sigma^2 = \frac{\sum^{m} f_i d_i^2}{N} = \frac{\sum^{m} f_i (x_i - \mu)^2}{N} \quad \text{(dados agrupados)}$$

EXEMPLO: a) $\sigma_1^2 = 0$ $\sigma_2^2 = 2/5=0,4$ $\sigma_3^2= 20/5=4$

$\sigma_4^2= 68/5=13,6$ $\sigma_5^2= 6/5=0,86$

b) Para os dados da Tabela 3: $\sigma^2= 232/50= 4,64$ anos^2

PROPRIEDADES:

P1:

$$\sigma^2 = \frac{\sum x_i^2}{N} - \mu^2 \quad \text{(dados não agrupados)}$$

$$\sigma^2 = \frac{\sum f_i x_i^2}{N} - \mu^2 \quad \text{(dados agrupados)}$$

Demonstração: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{\sum (x_i^2 - 2\, x_i\, \mu + \mu^2)}{N} =$

$$= \frac{\sum x_i^2 - 2\,\mu \sum x_i + \sum \mu^2}{N} = \frac{\sum x_i^2}{N} - 2\,\mu\,\frac{\sum x_i}{N} + \frac{N\,\mu^2}{N} =$$

$$= \frac{\sum x_i^2}{N} - 2\,\mu\,\mu + \mu^2 = \frac{\sum x_i^2}{N} - 2\,\mu^2 + \mu^2 = \frac{\sum x_i^2}{N} - \mu^2$$

IMPORTANTE: Esta é a fórmula prática para calcular a variância.

EXEMPLO: a) Para os dados do Conjunto 3:

x_i	x_i^2
2	4
4	16
5	25
6	36
8	64
Σ 25	145

$$\sigma^2 = \frac{\sum x_i^2}{N} - \mu^2 = \frac{145}{5} - 5^2 = 4$$

b) Para a Tabela 3:

Anos	f_i	x_i	$f_i x_i$	$f_i x_i^2$
0 |— 2	5	1	5	5
2 |— 4	15	3	45	135
4 |— 6	20	5	100	500
6 |— 8	5	7	35	245
8 |— 10	5	9	45	405
Σ	50		230	1299

$$\sigma^2 = \frac{\sum f_i x_i^2}{N} - \mu^2$$

$$= \frac{1299}{50} - 4{,}6^2$$

$$= 4{,}64 \text{ anos}^2$$

P2: A variância de uma constante é zero.

P3: A variância do produto de uma constante por uma variável é igual ao produto do quadrado da constante pela variância da variável.

P4: A variância da soma ou diferença de uma constante com uma variável é igual a variância da variável.

P5: A variância da soma ou diferença de duas variáveis independentes é igual a soma das variâncias

Erro relativo máximo cometido no cálculo da variância para dados agrupados em distribuição de frequência

$$\varepsilon = \frac{2\,h\,\mu}{\sigma^2 - 2\,h\,\mu} \qquad \text{para } \sigma^2 - 2\,h\,\mu > 0$$

onde h = amplitude de classe

μ = média da distribuição de frequência

σ^2 = variância da distribuição de frequência

OBSERVAÇÃO 1: A variância amostral s^2 é calculada pela mesma fórmula que a variância amostral, apenas substituindo-se N (tamanho da população) por n (tamanho da amostra).

É importante salientar que s^2 obtido desta maneira não se constitui uma boa estimativa para σ^2, isto é, se estivermos estimando a variância populacional, devemos utilizar uma correção, pois s^2 é um estimador viciado para σ^2. Neste caso, usamos $\hat{s}^2$ (variância corrigida) que é calculada por:

$$\hat{s}^2 = \frac{n}{n-1}\, s^2 \quad \text{(população infinita)}$$

$$\hat{s}^2 = \frac{N-1}{N}\,\frac{n}{n-1}\, s^2 \quad \text{(população finita)}$$

OBSERVAÇÃO 2: CÁLCULO DO PONTO MÉDIO DO INTERVALO ABERTO

Segundo Miller,P. - Rich Man, Poor Man- Thomas Y. Crowell Company,N.Y. (1971), quando o último intervalo não for limitado, o ponto médio será calculado por

$$y_i = y\,\frac{\alpha}{\alpha - 1} \qquad \text{onde} \qquad \alpha = \frac{c - d}{b - a}$$

y_i= ponto médio da última classe

y = limite inferior da classe aberta

a = log do limite inferior da penúltima classe

b = log do limite inferior da última classe

c = log da soma das frequências das duas últimas classes

d = log da frequência da última classe.

Desta forma, podemos calcular μ e σ^2 em distribuições de frequência com intervalo aberto na última classe.

2.1.5. DESVIO PADRÃO

Notação: σ (população)   s (amostra)

$$\sigma = \sqrt{\sigma^2}$$

O desvio padrão é também conhecido por afastamento padrão ou afastamento quadrático médio.

EXEMPLO: a) $\sigma_1 = 0$ $\sigma_2 = 0,63$ $\sigma_3 = 2$ $\sigma_4 = 3,68$ $\sigma_5 = 1,09$

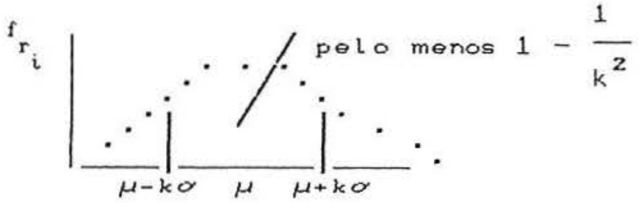
b) Para a Tabela 3: σ = 2,15 anos

IMPORTÂNCIA PRÁTICA DO DESVIO PADRÃO

I) TEOREMA DE TCHEBYSHEFF.

Dado um número k, maior ou igual a 1, e dado um conjunto de N observações de uma variável $X_1, X_2, \ldots, X_N$, é certo que pelo menos $\left(1 - \frac{1}{k^2}\right)$ dessas observações pertencerão ao intervalo de k desvios padrões antes e k desvios padrões além da média dessas observações.

A idéia envolvida neste resultado está no gráfico:



Um intervalo é constituido medindo-se uma distância $k\,\sigma$ para cada lado de μ.

Assim, calculando-se a fração $(1 - 1/k^2)$, vemos que pelo menos aquela fração do total N irá pertencer ao intervalo:

k	$1 - 1/_{k}2$
1	0
2	3/4
3	8/9

Isto é, para k = 2, observa-se que pelo menos 3/4 das observações estarão no intervalo $\mu \pm 2\,\sigma$. É claro que k = 0 não tem interesse prático, pois é incapaz de fornecer informações, já que diz que 0 observações estão no intervalo $\mu \pm \sigma$.

EXEMPLOS: (1)A média e a varância de 25 observações são 75 e 100, respectivamente. Usar o Teorema de Tchebysheff para descrever a distribuição.

Solução: a) Pelo menos 3/4 das 25 observações pertencem ao intervalo $\mu \pm 2\,\sigma = 75 \pm 2 \times 10 \Rightarrow [\ 55;\ 95\ ]$

b) Pelo menos 8/9 das 25 observações pertencem ao intervalo $\mu \pm 3\,\sigma \Rightarrow [\ 45;\ 105\ ]$

Podemos enfatizar o "pelo menos" do Teorema, por ser pessimista, quando aplicado a qualquer distribuição. Na maior parte dos casos, a fração de observações contidas neste intervalo irá exceder a $(1 - 1/\ k^2)$.

Se tivermos uma distribuição Normal, teremos que:

$\mu \pm \sigma$ contém 68 % das observações

$\mu \pm 2\,\sigma$ contém 95 % das observações

$\mu \pm 3\,\sigma$ contém praticamente todas as observações

(2) Realiza-se uma pesquisa a fim de avaliar o tempo necessário para a realização de certa operação manual numa fábrica. Esse tempo é medido para cada uma de 40 mulheres. A média e o desvio padrão foram 12,8 e 1,7, respectivamente. Descrever os dados.

Solução: $\mu \pm \sigma \Rightarrow [11{,}1\ ;\ 14{,}5]$

$\mu \pm 2\sigma \Rightarrow [9,4;\ 16,2]$

$\mu \pm 3\sigma \Rightarrow [7,7;\ 17,9]$

Supondo que seja Normal, é de se esperar que 68 %, 95 % e quase todas as observações estejam em cada um dos intervalos acima.

Se duvidarmos da normalidade, podemos usar o Teorema, sendo pessimistas, e dizer que pelo menos 3/4 das medidas, ou seja, 75 %, estão no intervalo 9,4 a 16,2 e pelo menos 8/9 das medidas, ou seja, 89 % estão no intervalo 7,7 a 17,9.

(3) Médias de alunos na disciplina MAT-201:

6,5	6,8	7,4	7,2	7,1	7,2	7,6	7,1	7,5	7,0
7,3	7,3	6,4	7,6	7,2	6,9	7,0	6,7	7,0	7,0
7,0	7,5	7,2	7,2	6,4	6,6	6,9	7,9	7,4	6,6

Calcule μ, σ, as frequências para cada intervalo $\mu \pm \sigma$, $\mu \pm 2\sigma$ e $\mu \pm 3\sigma$ e as f_{r_i}.

Solução: $\mu = 7,08$ $\sigma = 0,37$

k	$\mu \pm k\sigma$	f_i	f_{r_i}
1	6,71 a 7,45	19	0,63
2	6,34 a 7,82	29	0,97
3	5,97 a 8,19	30	1,00

Note que as f_{r_i} são próximas das % da Normal.

II) UMA AVALIAÇÃO DE σ por H

Podemos fazer uma estimativa grosseira de σ através de H, pois sabemos, pelo Teorema de Tchebysheff, que 3/4 do conjunto de observações está 2σ afastados da média. Consequentemente, a maior parte das observações pertence ao intervalo $\mu \pm 2\sigma$, de modo que a amplitude será aproximadamente igual a 4σ, isto é:

$$H \cong 4\sigma \quad \Rightarrow \quad \sigma \cong \frac{H}{4}$$

EXEMPLO: Calcular σ pela fórmula e avaliar σ pela aproximação indicada anteriormente: 85 70 60 90 81

$\mu = 77,2 \qquad \sigma^2 = 117,36 \Rightarrow \sigma = 10,8$

$$\sigma \cong \frac{H}{4} = \frac{90 - 60}{4} = \frac{30}{4} = 7,5$$

III) DETERMINAÇÃO DE h por σ

A escolha de h depende de fatores tais como o número de observações, a amplitude total, a precisão requerida no cálculo das estatísticas e o grau de condensação desejado. Para obtermos máxima precisão no cálculo das estatísticas numa distribuição de frequência, a amplitude de classe(h) não deveria ser, como regra, maior do que 1/4 do desvio padrão. Seguindo-se estritamente esta regra, porém, os dados muitas vezes não ficam adequadamente condensados para representação gráfica. É preferível aumentar os h para 1/3 ou 1/2 de σ e ignorar a perda de precisão no cálculo das estatísticas.

Como σ não é conhecido no momento em que se organiza a distribuição de frequência, poderemos estimá-lo a partir da relação que existe entre a amplitude total de variação e o desvio padrão associados com o número de observações.

A tabela a seguir mostra algumas dessas relações úteis na estimativa de σ:

Valores da relação amplitude/desvio padrão

N	H/σ	N	H/σ
20	3,7	200	5,5
30	4,1	300	5,8
50	4,5	400	5,9
70	4,8	500	6,1
100	5,0	700	6,3
150	5,3	1000	6,5

EXEMPLO: Numa fazenda foram registrados os pesos de 84 novilhos na fase final de engorda. Os pesos mínimo e máximo verificados foram 331 e 476 kg. Organize as classes.

Solução: H= 476 - 331 = 145 kg

Na tabela acima, a relação H/σ para N=84 é aproximadamente igual a 5,0. Estima-se o desvio padrão em:

$H/\sigma = 5,0 \Rightarrow 145/5,0 = 29 = \sigma$

Optou-se por 1/2 de σ para a amplitude de cada classe. Então $\sigma=29 \Rightarrow h = 29\ (1/2) - 14,5 \cong 15$.

Teremos então:

kg	f_i
331 |— 346	
346 |— 361	
361 |— 376	
376 |— 391	
391 |— 406	
406 |— 421	
421 |— 436	
436 |— 451	
451 |— 466	
466 |— 481	

IV) RELAÇÃO EMPÍRICA ENTRE DM (em relação a μ) e σ

Quando a distribuição for fracamente assimétrica, podemos usar a relação empírica:

$$DM = \frac{4}{5}\sigma$$

EXEMPLO: Consumo de energia elétrica

classes	f_i
5 |— 25	4
25 |— 45	6
45 |— 65	14
65 |— 85	26
85 |— 105	14
105 |— 125	8
125 |— 145	6
145 |— 165	2
Σ	80

$DM = 24,625$

$\sigma = 31,977$

$\Rightarrow DM = (4/5)31,977$
$= 25,582$

V) DESVIO REDUZIDO

O desvio padrão é utilizado no cálculo dos "desvios reduzidos" que é o afastamento entre o escore de um indivíduo e a média de seu grupo, dividido pelo desvio padrão do grupo:

$$z = \frac{x_i - \mu}{\sigma}$$

Estes desvios também recebem a denominação de "notas padrões", pois servem para comparar escores obtidos em provas diversas.

EXEMPLO:

Teste	μ	σ
Estatística	8	2
Cálculo	6	1,5

Se um aluno obteve 8,5 em Estatística e 8,0 em Cálculo, posso afirmar que o resultado foi superior em Estatística, considerando a variabilidade do grupo ?

$$z_{Est} = \frac{8,5 - 8}{2} = +0,25 \qquad z_{cál} = \frac{8,0 - 6}{1,5} = +1,33$$

Não podemos afirmar que o resultado do aluno foi superior em Estatística, se levarmos em consideração a variabilidade do grupo.

Demonstra-se que os desvios reduzidos formam uma escala com média zero e desvio padrão um.

DESVIO PADRÃO DE DISTRIBUIÇÃO COMBINADA

Muitas vezes, desejamos reunir numa só distribuição de frequência dados obtidos em dois ou mais grupos, dos quais já conhecemos as respectivas médias e desvios padrões.

Para tanto usa-se:

$$\sigma_G^2 = \frac{1}{N}\left[N_A\ (\ \sigma_A^2 + d_A^2\ )\ +\ N_B\ (\ \sigma_B^2 + d_B^2\ ) \right]$$

onde σ_G^2: variância da distribuição combinada ou variância geral

N : número total de observações

σ_A^2 : variância da 1ª distribuição

$d_A^2 = (\ \mu_A - \mu_G\ )^2$ sendo μ_A: média da 1ª distribuição

σ_B^2 : variância da 2ª distribuição

$d_B^2 = (\ \mu_B - \mu_G)^2$ sendo μ_B: média da 2ª distribuição

μ_G: média da distribuição combinada ou média geral

EXEMPLO:Seja um teste aplicado a duas turmas da mesma Escola

Turma	N	μ	σ
A	35	6,11	2,38
B	40	6,29	2,25

Média geral : $\mu_G = \dfrac{35 \times 6,11 + 40 \times 6,29}{75} = 6,25$

Variância geral:

$$\sigma_G^2 = \frac{1}{75} \left[35\ (5,66 + 0,01) + 40\ (5,06 + 0,0064) \right] = 5,35$$

$$\Rightarrow \quad \sigma_G = 2,31$$

Uma fórmula mais prática para se obter a variância de várias séries é:

$$\sigma_G^2 = \frac{\sum_{i=1}^{k} N_i\ \sigma_i^2 + \sum_{i=1}^{k} N_i\ \mu_i^2}{\sum_{i=1}^{k} N_i} - \mu_G^2$$

onde

$$\mu_G = \frac{\sum_{i=1}^{k} N_i\ \mu_i}{\sum_{i=1}^{k} N_i}$$

No exemplo:

$$\sigma_G^2 = \frac{35(2,38)^2 + 40(2,25)^2 + 35(6,11)^2 + 40(6,29)^2}{75} - 6,25^2 = 5,35$$

DECOMPOSIÇÃO DO DESVIO PADRÃO DE VÁRIAS SÉRIES

Consideremos k séries, constituidas, cada uma por N elementos, representados por x_{ij} com $i=1,2,\ldots,N$ e $j=1,2,\ldots,k$.

A média aritmética de todos os valores será dada por

$$\mu = \frac{\sum_{j=1}^{k} \sum_{i=1}^{N} x_{ij}}{k\ N}$$ e as diversas médias parciais

corresponderão a $\mu_1, \mu_2, \ldots, \mu_k$ onde

$$\mu_k = \frac{\sum_{i=1}^{N} x_{ik}}{N} .$$

Cada componente da série apresentará, em relação à média geral, um certo afastamento dado por $x_{ij} - \mu$. Por outro lado, os termos dos agrupamentos parciais apresentarão um desvio, em relação às médias das respectivas séries representados por $x_{ij} - \mu_j$.

Examinaremos então o caso dos desvios, por exemplo, relacionados à série de ordem j, integrante das k participantes no conjunto, sabendo-se que esta série tem N elementos.

O desvio de um termo desta série em relação à respectiva média parcial μ_j , será:

$$d_{1j} = x_{1j} - \mu_j \Rightarrow d_{1j} = x_{1j} - \mu_j + \mu - \mu \Rightarrow$$

$$d_{1j} - \mu = (x_{1j} - \mu_j) - \mu \Rightarrow (x_{1j} - \mu_j) - \mu = d_{1j} - \mu$$

Como $d_{1j} = x_{1j} - \mu_j$, podemos escrever:

$$x_{1j} - \mu_j - \mu = (x_{1j} - \mu_j) - \mu \quad \text{ou}$$

$$(x_{1j} - \mu) = (x_{1j} - \mu_j) + (\mu_j - \mu)$$

Elevando-se ao quadrado esta expressão e considerando a soma dos N termos da série, vem :

$$\sum_{i=1}^{N} (x_{ij} - \mu)^2 = \sum_{i=1}^{N} \left[(x_{ij} - \mu_j)^2 - 2(x_{ij} - \mu_j)(\mu_j - \mu) + (\mu_j - \mu)^2 \right]$$

$$= \sum_{i=1}^{N} (x_{ij} - \mu_j)^2 - 2(x_{ij} - \mu_j) \underbrace{\sum_{i=1}^{N} (\mu_j - \mu)}_{= 0} + \sum_{i=1}^{N} (\mu_j - \mu)^2$$

$$\sum_{i=1}^{N} (x_{ij} - \mu)^2 = \sum_{i=1}^{N} (x_{ij} - \mu_j)^2 + N (\mu_j - \mu)^2$$

Esta expressão corresponde à soma dos desvios em relação à série j; para as demais, o raciocínio é o mesmo Então, considerando-se as k séries, a soma dos quadrados dos dos desvios para todo o conjunto será:

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{N}(x_{ij}-\mu)^2}_{\text{soma dos quadrados dos desvios}} = \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}_{\text{soma dos quadrados dos desvios dentro das séries}} + \underbrace{N\sum_{j=1}^{k}(\mu_j-\mu)^2}_{\text{soma dos quadrados dos desvios entre as séries}}$$

Utilizando que

$$\sum (x_i-\mu)^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{N} \quad \text{e fazendo} \quad \sum_{j=1}^{k}\sum_{i=1}^{N} x_{ij} = \sum_{ij}^{kN} x_{ij}$$

obtém-se:

$$\underbrace{\sum_{ij}^{kN} x_{ij}^2 - \frac{(\sum^{kN} x_{ij})^2}{k\,N}}_{\text{SQ total}} = \underbrace{\left[\sum_{ij}^{kN} x_{ij}^2 - \sum_{j}^{k}\frac{(\sum^{N} x_{ij})^2}{N}\right]}_{\text{SQ dentro}} + \underbrace{\left[\sum_{j}^{k}\frac{(\sum^{N} x_{ij})^2}{N} - \frac{(\sum^{kN} x_{ij})^2}{k\,N}\right]}_{\text{SQ entre}}$$

Daí obtemos:

$$\sigma^2_{\text{Total}} = \frac{\sum_{ij}^{kN} x_{ij}^2 - \frac{(\sum^{kN} x_{ij})^2}{k\,N}}{k\,N} = \frac{\sum^{kN} x_{ij}^2}{k\,N} - \left(\frac{\sum^{kN} x_{ij}}{k\,N}\right)^2$$

$$\sigma^2_{\text{dentro}} = \frac{\sum^{kN} x_{ij}^2 - \sum^{K}\frac{(\sum^{N} x_{ij})^2}{N}}{k\,N} = \frac{\sum^{kN} x_{ij}^2}{kN} - \sum^{k}\frac{(\sum^{N} x_{ij})^2}{k\,N^2}$$

$$\sigma^2_{\text{entre}} = \frac{\sum^{k}\frac{(\sum^{N} x_{ij})^2}{N} - \frac{(\sum^{kN} x_{ij})^2}{kN}}{K} = \sum^{k}\frac{(\sum^{N} x_{ij})^2}{k\,N} - \frac{(\sum^{kN} x_{ij})^2}{k^2N}$$

EXEMPLO: Em determinada Escola, foi aplicado um teste a cinco grupos de alunos, obtendo-se os resultados:

									Σx_{ij}	Σx^2_{ij}	$(\Sigma x_{ij})^2$
A:	8	7	6	5	9	10	8	6	59	455	3481
B:	7	5	9	5	6	3	9	10	54	406	2916
C:	6	6	8	10	9	7	5	6	57	427	3249
D:	10	5	6	9	7	5	4	10	56	432	3136
E:	9	7	5	10	9	8	6	9	63	517	3969
								Σ	289	2237	16751

k = 5 N = 8 k N = 40

$$\sigma^2_{Total} = \frac{\sum_{i=1}^{5.8} x^2_{ij} - \dfrac{(\sum^{5.8} x_{ij})^2}{5.8}}{(5)(8)} = \frac{2237 - (289)^2/40}{40} = \frac{148,975}{40} = 3,72$$

$$\sigma^2_{dentro} = \frac{\sum^{5.8} x^2_{ij} - \sum^{5} \dfrac{(\sum^{8} x_{ij})^2}{8}}{(5)(8)} = \frac{2237 - 16751/8}{40} = \frac{143,125}{40} = 3,578$$

$$\sigma^2_{entre} = \frac{\sum^{5} \dfrac{(\sum^{8} x_{ij})^2}{8} - \dfrac{(\sum^{5.8} x_{ij})^2}{(8)(5)}}{5} = \frac{16751/8 - 289^2/40}{5} = \frac{5,85}{5} = 1,17$$

Os resultados indicam que os alunos apresentam maior homogeneidade entre grupos, uma vez que esta variância é menor.

Pelos cálculos, também podemos verificar que

$$148,987 = 143,125 + 5,85 ,$$

isto é, a soma de quadrados conjunta (total) é igual à soma dos quadrados dos desvios dentro e entre os grupos.

2.1.6. DIFERENÇA INTERQUARTIL

Notação: DI

É a diferença entre o 3º e o 1º quartil.

$DI = Q_3 - Q_1$

EXEMPLO: Para os dados da Tabela 2, temos Q_1= 167,25 cm e Q_3= 175,41 cm, então DI = 8,16 cm. Isto significa que 50%

das observações, correspondendo aos valores centrais, variam numa amplitude de 8,16 cm. Esquemàticamente:

$Q_1 = 167{,}25$ $\quad\quad$ $Q_3 = 175{,}41$

50%

2.1.7. DESVIO MEDIANO:

Notação : DQ

É a semi-diferença entre o $3^{\underline{o}}$ quartil e o $1^{\underline{o}}$ quartil. É também conhecido por desvio quartílico ou amplitude semiquartílica.

$$DQ = \frac{Q_3 - Q_1}{2}$$

O $1^{\underline{o}}$ quartil abrange cerca de 25% das ocorrências e o terceiro, 75%: assim, a diferença entre eles corresponderá a 50% dos valores que estão no centro da série.

Se a série for simétrica, devemos ter:
$Q_2 - Q_1 = Q_3 - Q_2$.

Tanto o desvio mediano como a diferença interquartil tem pouca sensibilidade, pois não consideramos o que ocorre entre os valores.

RELAÇÃO EMPIRICA ENTRE O DQ e σ.

Quando a distribuição se apresentar fracamente assimétrica, podemos usar a seguinte relação empírica:

$$DQ = \frac{2}{3}\sigma$$

2.2. MEDIDAS DE DISPERSÃO RELATIVA

As medidas de dispersão relativa permitem que se compare duas ou mais distribuições, mesmo que essas se refiram a diferentes fenômenos e sejam expressas em unidades de medida distintas.

2.2.1. VARIÂNCIA RELATIVA

Notação: γ

$$\gamma^2 = \frac{\sigma^2}{\mu^2}$$

EXEMPLO: Para a Tabela 3: $\gamma^2 = 4,64/(4,6)^2 = 0,2193$

2.2.2. COEFICIENTE DE VARIAÇÃO DE PEARSON

Notação: γ

$$\gamma = \frac{\sigma}{\mu}$$

EXEMPLO: Para os dados da Tabela 3: $\gamma = 2,15/4,6 = 0,4674$

IMPORTANTE: Se quisermos comparar duas ou mais variáveis que tenham médias diferentes ou diferentes unidades de medida, quanto à sua homogeneidade, isto é, grau de concentração em torno da média, devemos utilizar γ ou γ^2: a variável que tiver menor γ ou γ^2 será a mais homogênea.

Se as médias forem iguais, a que tiver menor σ ou σ^2 será a mais homogênea, isto é, não há necessidade de calcular γ ou γ^2.

2.2.3. DESVIO QUARTÍLICO REDUZIDO

Notação: DQ_r

É o resultado do quociente entre o desvio quartílico e a mediana.

$$DQ_r = \frac{Q_3 - Q_1}{2\ M_e}$$

EXEMPLO: Para os dados da Tabela 2, temos $Q_1 = 167,25$; $Q_3 = 175,41$ e $M_e = 172,47$, então : $DQ_r = 0,0002743$

2.2.4. COEFICIENTE DE VARIAÇÃO DE THORNDIKE

Notação: CV_T

$$CV_T = \frac{\sigma}{M_e}$$

EXEMPLO: O desvio padrão dos dados da Tabela 2 é 5,99, então CV_T = 5,99/172,47= 0,0347

2.2.5. COEFICIENTE QUARTÍLICO DE VARIAÇÃO

Notação: CV_Q

$$CV_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

2.3. EXERCÍCIOS

1. Os efeitos de duas drogas sobre a capacidade perceptual-motora foram os seguintes:

Grupo que recebeu a Droga 1: 8 6 7 9 8

Grupo que recebeu a Droga 2: 4 7 5 4 5 6 4

Qual das duas drogas produziu efeito mais homogêneo na capacidade perceptual-motora ?

2. Dois testes de aptidão (Compreensão Verbal e Aptidão Numérica) num grupo apresentaram os seguintes resultados:

$\bar{x}_{CV}$ = 28,86; s_{CV}= 12,55; $\bar{x}_{AN}$ = 19,89; s_{AN}= 7,37

Uma pessoa que obteve 35 pontos no teste de Aptidão Numérica e 42 pontos em Compreensão Verbal pode considerar que seu resultado foi superior neste último teste, levando em conta o resultado do grupo ?

3. Um teste de raciocínio foi aplicado a duas turmas U e V. A turma U tem 35 alunos e a turma V, 40. O quadro abaixo dá as médias e desvios padrões das duas turmas:

Turma	μ	σ
U	6,11	2,38
V	6,29	2,25

Calcule a média e o desvio padrão total.

4. Calcule o desvio padrão e o desvio médio dos seguintes conjuntos:

A: 1 6 7 8 9 10 14

B: 6 7 8 9 10

Qual das duas medidas de variabilidade é mais influenciada pelos valores extremos ?

5. Faça um confronto entre os resultados obtidos no vestibular de um Curso de Economia e os resultados obtidos ao final do ano letivo:

Vestibular: N= 100 $\mu = 6$ $M_e = 6,5$ $\sigma = 1,8$

Resultados ao final do ano letivo:

Médias	f_i
2	10
3	12
4	13
5	20
6	20
7	15
8	8
9	2
Σ	100

6. As estaturas de duas equipes (Verde x Amarela) são:

Verde: 1,60 1,66 1,66 1,68 1,70 1,75 1,82 1,83
1,83 1,85 1,85

Amarela: 1,71 1,72 1,73 1,74 1,75 1,76 1,76 1,76
1,77 1,77 1,78

Compare as equipes quanto à homogeneidade das alturas.

7. A média aritmética de dois números é 10 e a variância é 9. Determinar esses números.

8. Suponha que a média de pesos dos alunos da Turma X seja 44 kg com desvio padrão 5 kg. Deseja-se conhecer a média e o desvio padrão dos alunos de outra turma, sabendo que a relação entre as duas populações é $Y = 1,5\ X - 2$.

9. Sabendo que os principais produtos vendidos pela Granja Capivari são leite e ovos e que, durante o último mes, os valores médios diários foram:

Leite: $\mu = 150$ l; $\sigma = 3$ l Ovos: $\mu = 24$ dz; $\sigma = 1{,}2$ dz

Verificar onde houve maior regularidade nas vendas.

10. Idade dos visitantes de uma Exposição Científica

Anos	No. de visitantes
0 |— 10	6
10 |— 20	18
20 |— 30	11
30 |— 40	3
40 |— 50	1
50 |— 60	8
60 |— 70	4

Determine: a) a idade média dos visitantes; b) a classe onde se localiza a mediana; c) a moda de Czuber; d) o desvio médio; e) o desvio padrão.

11. Considerando que três variáveis apresentem os valores:

A	B	C
$N = 200$	$N = 50$	$\mu = 8$
$\Sigma f_i x_i = 4000$	$\Sigma f_i x_i = 500$	$\Sigma f_i x_i = 3200$
$\Sigma f_i (x_i - \mu)^2 = 5000$	$\Sigma f_i x_i^2 = 5450$	$\Sigma f_i x_i^2 = 32000$

Determine qual distribuição apresenta maior dispersão em torno da média.

12. Nas cidades Delta e Ômega foram obtidos os seguintes rendimentos de 10 indivíduos, em salários mínimos:

Delta: 5,50 8,20 6,00 5,50 4.70 4,70 7,20 3,60 5,50 68,10

Ômega: 8,00 15,00 9,50 11,25 12,00 14,50 15,00 8,75 14,50 10,50

a) Calcule e critique o rendimento médio para cada situação

b) Onde σ daria maior ? Verifique através dos cálculos.

13. Compare quanto à homogeneidade os grupos A e B de pessoas cujas pressões sanguíneas sistólicas (mm) foram:

A: 10 10 11 12 12 13 14 14 14 15

B: 7 7 8 9 12 13 13 16 17 23

14. Os dados de acidentes por mil homens/hora de uma amostra de 50 firmas apresentaram média 2,32 e desvio padrão 0,70. Uma amostra de 40 firmas de outro tipo de indústria tem média de 3,5 acidentes e desvio padrão 0,70 acidentes. Comparar a variabilidade dos dois tipos de indústrias quanto ao número de acidentes por mil homens/hora.

15. A medição da capacitância máxima de 20 condensadores fornecem os resultados (pF): 4,40 4,31 4,40 4,40 4,65 4,56 4,71 4,54 4,36 4,56 4,31 4,42 4,60 4,35 4,50 4,40 4,43 4,48 4,42 4,45

a) Calcule a média, o desvio padrão e o coeficiente de variação.

b) Organize uma distribuição de frequência com L_i= 4,31 e h=0,08 e recalcule as medidas anteriores.

16. Suponhamos que as seguintes medidas, em litros de óleo, para transformador foram obtidas a partir de duas companhias

Companhia	Observações					
A	19,8	19,9	20,1	20,2	20,4	19,6
B	19,5	20,1	19,9	20,2	20,3	20,0

Qual é a marca que deve ser preferida, supondo que o preço é o mesmo e que a decisão será tomada a favor da companhia que está mais próxima do volume nominal anunciado de 20 litros ?

17. Os valores abaixo referem-se aos diâmetros,em polegadas, de uma amostra de 60 canos fabricados por uma companhia:

0.738	0.729	0.743	0.740	0.736	0.741
0.728	0.737	0.736	0.735	0.724	0.733
0.745	0.736	0.742	0.740	0.728	0.738
0.733	0.730	0.732	0.730	0.739	0.734
0.735	0.732	0.735	0.727	0.734	0.732
0.732	0.737	0.731	0.746	0.735	0.729
0.737	0.735	0.732	0.735	0.744	0.740
0.735	0.738	0.731	0.739	0.729	0.727
0.742	0.736	0.736	0.741	0.739	0.736
0.725	0.734	0.733	0.735	0.734	0.730

Construa uma distribuição de frequência utilizando 6 classes e calcule a média aritmética, o desvio padrão e o coeficiente de variação.

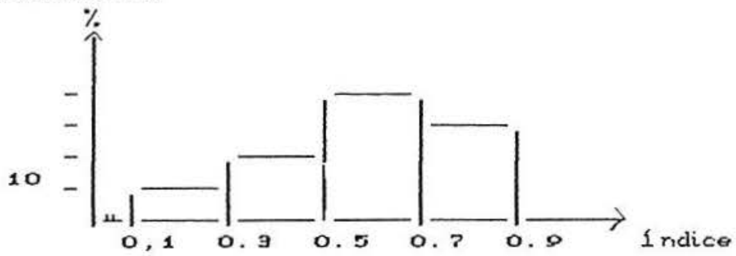
18. O principal objetivo do trabalho de Dina Araújo, "Taxonomia e Relações dos proganosauria da Bacia do Paraná", do Instituto de Geociências da UFRGS, é estabelecer decisões de ordem taxonômica com relação aos répteis mesossaurídeos, a partir de conjuntos de características apontados por vários autores, tais como medidas de crânio, pescoço, comprimento e largura dos dentes, comprimento e largura do úmero e fêmur. Consideremos os resultados referentes a medidas do crânio (cm) em exemplares ocorridos nas zonas A e B:

A					B				
7.1	8.1	7.5	7.1	6.7	6.0	5.8	5.8	5.3	4.9
6.7	8.9	7.1	7.2	7.6	5.8	6.1	6.6	4.7	6.4
7.9	7.8	9.2	7.1	9.6	5.8	6.2	5.0	6.5	5.2
5.1	7.9	7.2			6.2	6.5	5.2	5.5	4.1

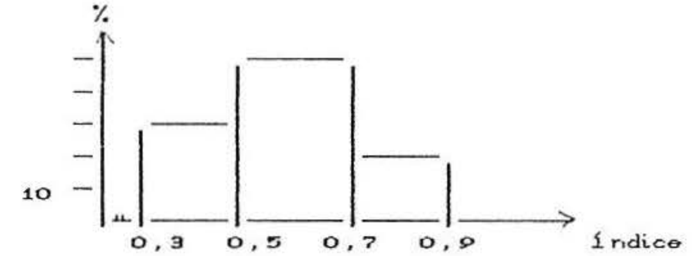
Compare as duas zonas quanto à homogeneidade.

19. Em "Notas e Estudos", de maio de 1970, da então Escola de Geologia da UFRGS, consta um trabalho sobre "Sedimentos praiais de Santa Catarina- Laguna e Imbituba", cuja finalidade é a caracterização dos sedimentos praiais deste litoral quanto aos índices de esfericidade, arredondamento e textura superficial que constituem propriedades morfoscópicas. Uma amostra com 20 observações de areia de praia em certo ponto apresentou os resultados:

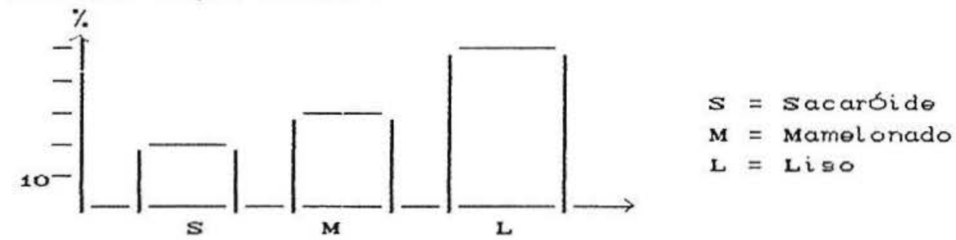
a) arredondamento



b) esfericidade



c) esfericidade superficial:



a)Compare as três características quanto à ocorrência mais frequente.

b) Diga em relação a qual característica quantitativa há maior homogeneidade.

20. A resistência de um adesivo é determinada pela força necessária para separar itens ligados pelo adesivo. Após serem ensaiados 16 itens para os dois tipos de adesivos S_1 e S_2, foram obtidos os seguintes valores:

Tipo de adesivo	Força de ruptura (kgf)							
S_1	10,2	5,8	4,9	16,1	15,0	9,4	4,8	10,1
	14,6	19,7	5,0	4,7	16,8	4,5	4,0	16,5
S_2	7,6	13,7	7,0	12,8	11,8	13,7	14,8	10,4
	7,0	10,1	6,8	10,0	8,6	11,2	8,3	10,6

Compare S_1 e S_2 quanto à homogeneidade da força de ruptura.

21. A qualidade de um produto é definida através de duas características: número da dureza Brinell (NDB) e diâmetro.

As especificações dessas características são:

NDB: 250 ± 5 Diâmetro: $1{,}000 \pm 0{,}002$ pol

Os seguintes valores da dureza Brinell foram encontrados:

248	250	249	252	253	249	247	249	250	251
250	249	248	250	251	249	245	246	249	254

Foram também registrados os valores do diâmetro:

1,0010	0,9998	0,9980	1,0000	1,0020	0,9999	1,0010
1,0013	1,0015	1,0020	1,0009	1,0009	1,0009	1,0011
0,9996	1,0009	1,0019	0,9997	0,9990	1,0009	

a) Qual é o percentual de produtos fora da especificação, para cada característica ?

b) Relativamente a qual característica existe maior dispersão em torno da média.

c) A perda causada por dureza Brinell não-conforme é 20 u.m. e a causada pelo tamanho do diâmetro não-conforme é 30 u.m. Qual foi a perda total por desvios dos valores nominais para essa amostra?

22. Um fabricante de trocadores de calor exige que o espaçamento na placa seja 0,25 $\pm$ 0,01 pol. O estatístico responsável pelo controle de qualidade colheu amostras de 25 trocadores de calor e mediu aleatoriamente o espaçamento entre duas placas em cada trocador. Foram registradas as seguintes medidas:

0,251	0,248	0,244	0,246	0,249	0,248	0,249
0,250	0,254	0,253	0,241	0,243	0,251	0,256
0,257	0,251	0,240	0,249	0,258	0,259	0,249
0,245	0,253	0,251	0,250			

a) A perda devido a um trocador não-conforme é 50 u.m. (custo do ajuste do espaçamento). Qual foi a perda total nessa amostra ? b) Analise os dados quanto à tendência central e variabilidade.

23. A tolerância em uso numa certa folga do produto é 15 μm ou menos. Produtos que não correspondem a uma certa especificação são escrapeados. Há dois tipos de máquinas, A_1 e A_2, que podem ser usadas para produzir esse ítem. O quadro a seguir contém dados sobre 15 peças produzidas pelos dois tipos de máquinas:

Folga (μm)															
A_1	3	7	4	13	8	4	7	3	6	7	5	9	4	4	3
A_2	8	9	11	9	7	9	11	6	7	14	14	11	11	9	7

Qual das duas máquinas produz ítens mais homogêneos quanto à folga ?

24. Um fornecedor de eletrônica automotiva utiliza transdutor interferométrico a laser cujos sinais são processados por microcomputador, para assegurar que a válvula esférica de regulagem da injeção de combustível funcione, durante um bilhão de ciclos - ou 400 000 milhas de percurso. O limite superior da tolerância da circularidade entre a esfera e a base do injetor é 1 μm, enquanto que o limite inferior é 0,5 μm e o valor nominal é 0,8 μm. O custo de um injetor não-conforme quando a circularidade excede 1 μm é de 107 u.m., enquanto que o custo de injetor não-conforme quando a circularidade for menor do que 0,5 μm é de 160 u.m. Foram registradas as seguintes tolerâncias (em micrometros):

0,60 0,90 0,78 0,87 0,78 0,70 0,91 0,98 0,90 0,68
0,55 0,86 0,76 0,86 0,91 0,59 0,57 0,50 0,87 0,92
0,81 0,86 0,50 0,85 0,69

a) Qual foi o custo total dos injetores não-conformes nesta amostra ? b) Faça uma análise descritiva dos dados.

III. CARACTERÍSTICAS DE ASSIMETRIA E CURTOSE

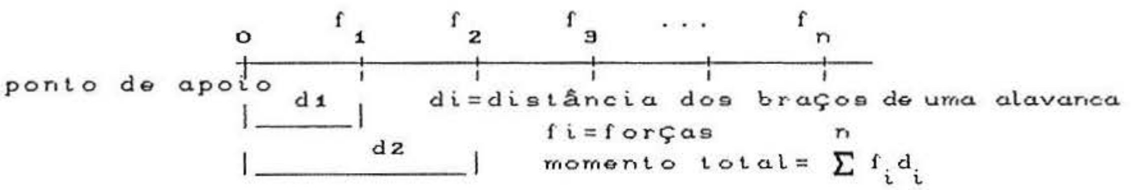
As medidas de assimetria e curtose complementam o quadro das estatísticas descritivas que proporcionam, juntamente com as medidas de tendência central e variabilidade, a descrição da variável em estudo.

As características mais importantes quanto à forma da distribuição da varivel são o grau de deformação ou assimetria e o grau de achatamento,afilamento ou curtose. Para estudar as medidas de assimetria e de curtose é necessário conhecer certas quantidades conhecidas por MOMENTOS.

1. MOMENTOS

Momentos são valores numéricos extraídos de uma variável. Através dos momentos pode-se determinar certos parâmetros como μ , σ^2 e as medidas de assimetria e curtose.

A concepção de momento em Estatística é semelhante à Física. Por exemplo, na Mecânica: Momento de uma força em relação a um ponto é o produto da potência dessa força pela distância do ponto de apoio.



Os momentos podem ser calculados em relação:

1º) a uma origem zero;

2º) à média aritmética;

3º) a um ponto qualquer, ou seja, uma origem arbitrária.

Os dois primeiros casos são situações particulares do 3º.

1.1. MOMENTO ORDINÁRIO OU MOMENTO NATURAL DE ORDEM r

Notação: μ'_r (população) m'_r(amostra) onde $r=0,1,2,\ldots$

Definição:

$$\mu'_r = \frac{\sum_{i=2}^{N} x_i^r}{N} \quad \text{(dados não agrupados)}$$

$$\mu'_r = \frac{\sum_{i=2}^{m} f_i x_i^r}{N} \quad \text{(dados agrupados)}$$

Para r=0 :

$$\mu'_o = \frac{\sum_{i=2}^{m} f_i x_i^o}{N} = \frac{\sum f_i}{N} = \frac{N}{N} = 1$$

Para r=1:

$$\mu'_1 = \frac{\sum_{i=2}^{m} f_i x_i^1}{N} = \frac{\sum f_i x_i}{N} = \mu$$

Para r=2:

$$\mu'_2 = \frac{\sum_{i=2}^{m} f_i x_i^2}{N} \quad \text{(quadrado da média quadrática)}$$

Para r=3:

$$\mu'_3 = \frac{\sum_{i=2}^{m} f_i x_i^3}{N} \quad \text{(cubo da média cúbica)}$$

OBSERVAÇÃO:

$$\sigma^2 = \mu'_2 - (\mu'_1)^2$$

EXEMPLO 1: Calcular o momento ordinário de $1^{\underline{a}}$, $2^{\underline{a}}$, $3^{\underline{a}}$ e $4^{\underline{a}}$ ordens da seguinte distribuição de frequência

Salários dos funcionários da Empresa TIMBAVER

salários mínimos	f_i	x_i	$f_i x_i$	$f_i x_i^2$	$f_i x_i^3$	$f_i x_i^4$
1 |— 3	2	2	4	8	16	32
2 |— 5	4	4	16	64	256	1024
3 |— 7	8	6	48	288	1728	10368
4 |— 9	4	8	32	256	2048	16384
5 |— 11	2	10	20	200	2000	20000
Σ	20		120	816	6048	47808

$\mu_1 = 120/20 = 6 = \mu$ $\qquad$ $\mu'_2 = 816/20 = 40,8$

$\mu'_3 = 6048/20 = 302,4$ $\qquad$ $\mu'_4 = 47808/20 = 2390,4$

1.2. MOMENTO CENTRAL OU MOMENTO DE ORDEM r CENTRADO NA MÉDIA

Notação: μ_r (população) $\quad$ m_r (amostra)

Definição:

$$\mu_r = \frac{\sum_{i=1}^{N} (x_i - \mu)^r}{N} \qquad \text{(dados não agrupados)}$$

$$\mu_r = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^r}{N} \qquad \text{(dados agrupados)}$$

Para r=0:

$$\mu_o = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^o}{N} = \frac{\Sigma f_i}{N} = 1$$

Para r=1:

$$\mu_1 = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^1}{N} = \frac{\Sigma f_i \ d_i}{N} = 0$$

Para r=2:

$$\mu_2 = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^2}{N} = \sigma^2$$

OBSERVAÇÃO: Como

$$\sigma^2 = \frac{\Sigma f_i x_i^2}{N} - \mu^2 = \mu_2' - (\mu_1')^2$$

temos que: $\mu_2 = \mu_2' - (\mu_1')^2$.

(momento central de ordem 2 em função dos momentos ordinários)

Para r=3:

$$\mu_3 = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^3}{N} \quad \text{(medida de assimetria)}$$

que desenvolvendo obtém-se:

$$\mu_3 = \frac{\Sigma f_i x_i^3}{N} - 3 \frac{\Sigma f_i x_i}{N} \frac{\Sigma f_i x_i^2}{N} + 2 \left[\frac{\Sigma f_i x_i}{N}\right]^2 \left[\frac{\Sigma f_i x_i}{N}\right] =$$

$$= \mu_3' - 3 \mu_1' \mu_2' + 2 (\mu_1')^3$$

Para r=4:

$$\mu_4 = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^4}{N} \quad \text{(medida de curtose)}$$

que é:

$$\mu_4 = \frac{\Sigma f_i x_i^4}{N} - 4 \frac{\Sigma f_i x_i^3}{N} \frac{\Sigma f_i x_i}{N} + 6 \left[\frac{\Sigma f_i x_i}{N}\right]^2 \frac{\Sigma f_i x_i^2}{N} -$$

$$3 \left[\frac{\Sigma f_i x_i}{N}\right]^4 = \mu_4' - 4 \mu_1' \mu_3' + 6 (\mu_1')^2 \mu_2' - 3 (\mu_1')^4$$

EXEMPLO: Calcular os momentos centrais de 2ª, 3ª e 4ª ordens e a variância do Exemplo 1.

$\mu_2 = 40,8-(6)^2 = 4,8 = \sigma^2$ $\quad\quad$ $\mu_3 = 302,4 - 3(6)(40,8)+2(6)^3=0$

$\mu_4 = 2390,4 - 4(6)(302,4) + 6(6)^2(40,8) - 3(6)^4 = 57,6$

1.3. MOMENTO DE ORDEM r EM RELAÇÃO A UMA ORIGEM QUALQUER x_o

Notação: ${}_{x_o}\mu_r$ (população) $\quad$ ${}_{x_o}m_r$ (amostra)

Definição:

$$\mu_r = \frac{\sum_{i=1}^{N} (x_i - x_o)^r}{N} \quad \text{(dados não agrupados)}$$

$$\mu_r = \frac{\sum_{i=1}^{m} f_i (x_i - x_o)^r}{N} \quad \text{(dados agrupados)}$$

É o momento genérico de **1.1** e **1.2.**

Podemos estabelecer relações entre os momentos centrados na média e os momentos centrados numa origem arbitrária x_o .

Para r=2: $\mu_2 = {}_{x_o}\mu_2 - ({}_{x_o}\mu_1)^2$

Para r=3: $\mu_3 = {}_{x_o}\mu_3 - 3\,({}_{x_o}\mu_1)\,({}_{x_o}\mu_2) + 2\,({}_{x_o}\mu_1)^3$

Para r=4:

$$\mu_4 = {}_{x_o}\mu_4 - 4\,({}_{x_o}\mu_1)\,({}_{x_o}\mu_3) + 6\,({}_{x_o}\mu_1)^2\,({}_{x_o}\mu_2) - 3\,({}_{x_o}\mu_1)^4$$

EXEMPLO: Calcular os momentos de $1^{\underline{a}}$, $2^{\underline{a}}$, $3^{\underline{a}}$ e $4^{\underline{a}}$ ordens em relação à origem $x_o = 4$, com os dados do Exemplo1. Utilizando esses momentos, calcular os momentos centrados na média.

x_i	f_i	(x_i-4)	$f_i(x_i-4)$	$f_i(x_i-4)^2$	$f_i(x_i-4)^3$	$f_i(x_i-4)^4$
2	2	-2	-4	8	-16	32
4	4	0	0	0	0	0
6	8	2	16	32	64	128
8	4	4	16	64	256	1024
10	2	6	12	72	432	2592
Σ	20	/	40	176	736	3776

${}_{x_o}\mu_1 = 40/20 = 2$ ${}_{x_o}\mu_2 = 176/20 = 8{,}8$

$$x_o\mu_3 = 736/20 = 36,8 \qquad x_o\mu_4 = 3776/20 = 188,8$$

$$\mu_2 = 8,8 - (2)^2 = 4,8 \qquad \mu_3 = 36,8 - 3(2)(8,8) + 2(2)^3 = 0$$

$$\mu_4 = 188,8 - 4(2)(36,8) + 6(2)^2(8,8) - 3(2)^4 = 57,6$$

CORREÇÃO DE SHEPPARD PARA OS MOMENTOS

Os 2º e 4º momentos centrados são corrigidos, para distribuições de frequência por intervalo, através das expressões:

$$\mu_{2(corr)} = \mu_2 - \frac{1}{12} h^2$$

$$\mu_{4(corr)} = \mu_4 - \frac{h^2 \mu_2}{2} + \frac{7 h^4}{240}$$

Os momentos de 1ª e 3ª ordem não são corrigidos.

1.4. MOMENTOS CENTRAIS RELATIVOS OU REDUZIDOS

Dividindo-se os momentos centrais por unidades de desvios padrões, tem-se os momentos relativos ou reduzidos.

$$\alpha_r = \frac{\mu_r}{\sigma^r}$$

Para r=0:

$$\alpha_o = \frac{\mu_o}{\sigma^o} = 1$$

Para r=1:

$$\alpha_1 = \frac{\mu_1}{\sigma^1} = 0$$

Para r=2:

$$\alpha_2 = \frac{\mu_2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

Para r=3:

$$\alpha_3 = \frac{\mu_3}{\sigma^3} \quad \text{(medida de assimetria)}$$

Para r=4:

$$\alpha_4 = \frac{\mu_4}{\sigma^4} \quad \text{(medida de curtose)}$$

EXEMPLO: Retomando os dados do Exemplo 1, temos $\alpha_3 = 0$ e $\alpha_4 = 57,6/(4,8)^2 = 2,5$.

2. COEFICIENTES DE ASSIMETRIA E CURTOSE

2.1. COEFICIENTE DE ASSIMETRIA DE PEARSON

Notação: β_1

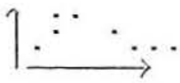
O quadrado de α_3 é o coeficiente de assimetria de Pearson:

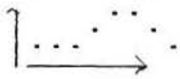
$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{(\sigma^3)^2} = \frac{\mu_3^2}{\sigma^6} = \frac{\mu_3^2}{\mu_2^3}$$

pois $\sigma^2 = \mu_2 \Rightarrow \sigma^6 = \mu_2^3$

Temos: $\beta_1 = 0 \Rightarrow$ curva simétrica

$\beta_1 \neq 0 \Rightarrow$ curva assimétrica :

$\mu_3 > 0 \Rightarrow$ assimetria positiva 

$\mu_3 < 0 \Rightarrow$ assimetria negativa 

Muitas vezes usa-se α_3 como coeficiente de assimetria.

EXEMPLO:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)^2}{(4,8)^3} = 0$$

2.2. COEFICIENTE MOMENTO DE CURTOSE

Notação: β_2

$$\beta_2 = \alpha_4 = \frac{\mu_4}{\sigma^4}$$

Se: $\beta_2 < 3 \Rightarrow$ curva achatada (platicúrtica)
$\beta_2 = 3 \Rightarrow$ mesocúrtica
$\beta_2 > 3 \Rightarrow$ curva ponteaguda (leptocúrtica)

A curtose indica até que ponto a curva de frequências de uma distribuição se apresenta mais afilada ou mais achatada do que uma curva Normal.

EXEMPLO: $\beta_2 = \dfrac{\mu_4}{\sigma^4} = \dfrac{57,6}{(4,8)^2} = 2,5$

2.3. COEFICIENTE C_2

A curtose pode ser igualmente medida pelo coeficiente C_2:

$$C_2 = \beta_2 - 3$$

Se: $C_2 < 0 \Rightarrow$ platicúrtica
$C_2 = 0 \Rightarrow$ mesocúrtica
$C_2 > 0 \Rightarrow$ leptocúrtica

EXEMPLO: $C_2 = -0,5 \Rightarrow$ curva platicúrtica

3. OUTRAS MEDIDAS DE ASSIMETRIA

3.1. MÉTODO DE COMPARAÇÃO ENTRE MEDIDAS DE TENDÊNCIA CENTRAL

Só vale para distribuição unimodal:

$\mu > M_o \Rightarrow$ assimetria positiva

$\mu = M_o \Rightarrow$ simétrica

$\mu < M_o \Rightarrow$ assimetria negativa

3.2. COEFICIENTES (ÍNDICES) DE PEARSON

3.2.1. PRIMEIRO COEFICIENTE DE ENVIESAMENTO (ASSIMETRIA) DE PEARSON

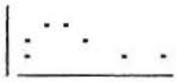
$$e_1 = \frac{\mu - M_o}{\sigma}$$

3.2.2. SEGUNDO COEFICIENTE DE ASSIMETRIA DE PEARSON

$$e_2 = \frac{3\ (\mu - M_e)}{\sigma} \quad \text{onde} \quad -3 \leq e_2 \leq 3$$

EXEMPLO: Admitindo-se uma distribuição com $\mu = 95$; $M_e = 90$ e $\sigma = 33,33$, calcular e_2 e interpretar.

$$e_2 = \frac{3\ (95 - 90)}{33,33} = 0,45 \quad \Rightarrow \text{assimetria positiva}$$



3.3. COEFICIENTE QUARTIL DE ASSIMETRIA

$$e_Q = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} \quad \Rightarrow$$

$$e_Q = \frac{Q_3 - 2\ M_e + Q_1}{Q_3 - Q_1} \quad \text{onde} \quad -1 \leq e_Q \leq 1$$

3.4. COEFICIENTE DE ASSIMETRIA ENTRE OS PERCENTIS 10 E 90

$$e_C = \frac{(C_{90} - M_e) - (M_e - C_{10})}{(C_{90} - M_e) + (M_e - C_{10})} \Rightarrow$$

$$e_C = \frac{C_{90} - 2\,M_e + C_{10}}{C_{90} - C_{10}}$$

3.5. COEFICIENTE MOMENTO DE ASSIMETRIA

$$e_{M_1} = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\,\beta_2 - 6\,\beta_1 - 9)} \quad \text{onde} \quad \beta_1 = \frac{\mu^2_3}{\mu^3_2} \quad \text{e} \quad \beta_2 = \frac{\mu_4}{\mu^2_2}$$

Constatou-se que este coeficiente não diferencia situações com mesmo β_1 e β_2 . A diferença na assimetria será acusada por α_3.

4. OUTRA MEDIDA DE CURTOSE:
COEFICIENTE PERCENTÍLICO DE CURTOSE

$$K = \frac{D_Q}{C_{90} - C_{10}}$$

onde

$$D_Q = \frac{Q_3 - Q_1}{2}$$

Se: $K > 0{,}263 \Rightarrow$ platicúrtica
$K = 0{,}263 \Rightarrow$ mesocúrtica
$K < 0{,}263 \Rightarrow$ leptocúrtica

5. EXERCÍCIOS

1. Compare as seguintes distribuições de frequência quanto à tendência central (μ), variabilidade (σ e γ), assimetria (α_3) e curtose (α_4). Faça um histograma para cada uma.

classes	f_{1i}	f_{2i}	f_{3i}	f_{4i}	f_{5i}	f_{6i}	f_{7i}	f_{8i}	f_{9i}	f_{10i}
$0 \vdash 2$	5	5	10	20	20	5	30	2	2	30
$2 \vdash 4$	10	15	10	8	10	5	10	3	8	8
$4 \vdash 6$	20	10	10	1	10	10	5	5	30	8
$6 \vdash 8$	10	15	10	8	5	10	3	10	8	2
$8 \vdash 10$	5	5	10	20	5	20	2	30	2	2
Σ	50	50	50	50	50	50	50	50	50	50

distribuições / medidas	1	2	3	4	5	6	7	8	9	10
μ										
σ										
γ										
α_3										
α_4										
Interpretação										
Gráfico										

2. Porcentagem de bactérias por cm^2 em 100 amostras de determinado produto

%	f_i
$0 \vdash 0,2$	7
$0,2 \vdash 0,4$	25
$0,4 \vdash 0,6$	36
$0,6 \vdash 0,8$	25
$0,8 \vdash 1,0$	7
Σ	100

Calcular e interpretar:

a) coeficiente percentílico de curtose

b) coeficiente quartil de assimetria

c) primeiro coeficiente de assimetria de Pearson

d) segundo coeficiente de assimetria de Pearson

e) coeficiente de assimetria entre os percentis 10 e 90.

3. Num teste de habilitação a média foi 61, a moda 57,5 e o desvio padrão 1,68. Indicar a assimetria através do primeiro coeficiente de Pearson.

4. Estudar os tempos de execução de uma tarefa para os operários A e B quanto à média, variância, desvio padrão, coeficiente de variação, beta 1 e beta 2. Discutir os resultados.

A: 74 75 78 78 72 77 79 78 81 76 72 72

B: 86 84 80 80 88 89 85 86 82 82 84 79 80 82 76

5. Idade dos tenistas do Clube Saúde

Anos	f_i
20 ⊢ 26	10
26 ⊢ 32	20
32 ⊢ 38	30
38 ⊢ 44	16
44 ⊢ 50	4
Σ	80

Determinar e, sempre que possível, interpretar.

a) mediana ; b) primeiro quartil; c) terceiro quartil;

d) momento natural de $1^{\underline{a}}$, $2^{\underline{a}}$, $3^{\underline{a}}$ e $4^{\underline{a}}$ ordens;

e) momento central de $2^{\underline{a}}$, $3^{\underline{a}}$ e $4^{\underline{a}}$ ordens;

f) variância absoluta;

g) desvio padrão;

h) coeficiente de variação;

i) coeficiente de assimetria de Pearson;

j) coeficiente momento de curtose;

k) primeiro e segundo coeficientes de assimetria de Pearson;

l) coeficiente quartil de assimetria.

IV. DISTRIBUIÇÃO DE VARIÁVEL BIDIMENSIONAL. NOÇÕES DE AJUSTAMENTO.

1. DISTRIBUIÇÃO DE VARIÁVEL BIDIMENSIONAL

Muitas vezes, temos interesse em estudar duas ou mais variáveis simultaneamente. Vamos explorar algumas idéias sobre a descrição tabular de duas variáveis.

EXEMPLO 1: Um levantamento junto aos alunos de MAT 271, em 92/2, forneceu os seguintes dados:

Aluno	Sexo	Idade
1	M	19
2	M	20
3	M	19
4	F	18
5	M	18
6	M	19
7	M	20
8	M	20
9	F	20
10	F	20

1.1. DISTRIBUIÇÃO BIDIMENSIONAL OU DISTRIBUIÇÃO CONJUNTA

Dadas as observações de duas variáveis, podemos organizar uma tabela de dupla entrada chamada de distribuição conjunta ou distribuição bidimensional.

EXEMPLO: Para os dados do Exemplo 1, construimos a Distribuição conjunta de frequências absolutas simples:

Tabela 1. Distribuição dos Alunos de MAT 271 segundo o Sexo e a Idade

Sexo \ Idade	18	19	20	Σ
M	1	3	3	7
F	1	0	2	3
Σ	2	3	5	10

Nesta situação, a variável Sexo é a variável x_i com $i=1,2$ e a variável Idade é a variável y_j para $j = 1, 2,3$. Dentro da tabela, estão as frequências f_{ij} . Assim, f_{23} é a frequência da segunda linha e terceira coluna, isto é, $f_{23}=2$ significa que 2 alunos são do sexo feminino e tem 20 anos.

1.2. DISTRIBUIÇÃO MARGINAL

A partir da distribuição conjunta, podemos extrair as distribuições marginais , ou seja,as distribuições de frequência para cada variável.

EXEMPLO: Considerando a Tabela 1, obtemos:

a) Distribuição Marginal de X

Tabela 2. Distribuição dos Alunos de MAT 271 segundo o Sexo

Sexo	f_i
M	7
F	3
Σ	10

b) Distribuição Marginal de Y

Tabela 3. Distribuição dos Alunos de MAT 271 segundo a Idade

Idade	f_j
18	2
19	3
20	5
Σ	10

1.3. DISTRIBUIÇÃO CONDICIONAL

Com base na Distribuição Conjunta de frequências absolutas, podemos construir as Distribuições Condicionais de frequência absoluta de X para um dado valor de Y ou de Y para um dado valor de X.

EXEMPLO: Para os dados da Tabela 1:

a) Distribuição Condicional de $X / Y_1= 18$

Tabela 4. Distribuição dos Alunos de MAT 271 com 18 anos de idade segundo o sexo

Sexo	f_i
M	1
F	1
Σ	2

b) Distribuição Condicional de $X / Y_2 = 19$

Tabela 5. Distribuição dos Alunos de MAT 271 com 19 anos de idade segundo o sexo

Sexo	f_i
M	3
F	0
Σ	3

c) Distribuição Condicional de $X / Y_3 = 20$

Tabela 6. Distribuição dos Alunos de MAT 271 com 20 anos de idade segundo o sexo

Sexo	f_i
M	3
F	2
Σ	5

d) Distribuição Condicional de $Y / X_1 = M$

Tabela 7. Distribuição dos Alunos de MAT 271 do Sexo Masculino segundo a Idade

Idade	f_j
18	1
19	3
20	3
Σ	7

e) Distribuição Condicional de $Y / X_2 = F$

Tabela 8. Distribuição dos Alunos de MAT 271 do Sexo Feminino segundo a Idade

Idade	f_j
18	1
19	0
20	2
Σ	3

1.4. DISTRIBUIÇÃO CONJUNTA DE FREQUÊNCIA RELATIVA

EXEMPLO:

Tabela 9. Distribuição dos Alunos de MAT 271 segundo o Sexo e a Idade

Sexo \ Idade	18	19	20	Σ
M	0,1	0,3	0,3	0,7
F	0,1	0	0,2	0,3
Σ	0,2	0,3	0,5	1,0

1.5. DISTRIBUIÇÃO MARGINAL DE FREQUÊNCIA RELATIVA

EXEMPLO: a) M(X)

Sexo	f_i
M	0,7
F	0,3
Σ	1,0

b) M(Y)

Idade	f_j
18	0,2
19	0,3
20	0,5
Σ	1,0

1.6. DISTRIBUIÇÃO CONDICIONAL DE FREQUÊNCIA RELATIVA

As Distribuições Condicionais das frequências relativas podem ser obtidas através da Tabela 1 ou das Tabelas 4 a 8 se as frequências absolutas estiverem disponíveis; ou através da Tabela 9 se as frequências relativas estiverem disponíveis.

De modo geral, a frequência relativa condicional de Y_j da X_i é dada por

$$f_r\ (Y_j / X_i) = \frac{f_{ij}}{f_{i.}}$$

Assim,

$$f_r(Y_1 / X_1) = \frac{f_{11}}{f_{1.}} = \frac{1}{7} \qquad f_r(Y_1/ X_2) = \frac{f_{21}}{f_{2.}} = \frac{1}{3}$$

$$f_r(Y_2 / X_1) = \frac{f_{12}}{f_{1.}} = \frac{3}{7} \qquad f_r(Y_2/ X_2) = \frac{f_{22}}{f_{2.}} = \frac{0}{3}$$

$$f_r(Y_3 / X_1) = \frac{f_{13}}{f_{1.}} = \frac{3}{7} \qquad f_r(Y_3/ X_2) = \frac{f_{23}}{f_{2.}} = \frac{2}{3}$$

Tabela 10. Distribuição Condicional de $Y / X_1 = M$

Y_j	$f_r(Y_j / X_1)$
18	1/7
19	3/7
20	3/7
Σ	1

De modo análogo :

$$f_r (X_i / Y_j) = \frac{f_{ij}}{f_{.j}}$$

é a frequência relativa condicional de X_i dado Y_j .

Se usarmos a Tabela 9, teremos:

$$f_r (X_i / Y_j) = \frac{fr_{ij}}{fr_{.j}} \quad \text{e} \quad f_r (y_j / X_i) = \frac{fr_{ij}}{fr_{i.}} .$$

Assim,

$$f_r(X_1/Y_3) = \frac{fr_{13}}{fr_{.3}} = \frac{0,3}{0,5} = \frac{3}{5} \qquad f_r(X_2/Y_3) = \frac{fr_{23}}{fr_{.3}} = \frac{0,2}{0,5} = \frac{2}{5}$$

Tabela 11. Distribuição Condicional de $X / Y_3 = 20$

x_i	$f_r (x_i / y_3)$
M	3/5
F	2/5
Σ	1

2. INDEPENDÊNCIA DE DUAS VARIÁVEIS

Duas variáveis X e Y são independentes se

$$f_{ij} = \frac{f_{i.} \cdot f_{.j}}{N}$$

EXEMPLO: Consideremos as seguintes distribuições bidimensionais:

a) Tabela 12.

X \ Y	0	1	2	Σ
0	10	20	20	50
1	4	8	8	20
2	6	12	12	30
Σ	20	40	40	100

X e Y são independentes pois:

$$f_{11} = \frac{f_{1.} \cdot f_{.1}}{N} = \frac{50.20}{100} = 10$$

$$f_{12} = \frac{f_{1.} \cdot f_{.2}}{N} = \frac{50.40}{100} = 20$$

e assim por diante.

b) Tabela 13.

X \ Y	0	1	2	Σ
0	5	10	20	35
1	8	12	15	35
2	6	8	16	30
Σ	19	30	51	100

X e Y não são independentes pois pelo menos uma das

$$f_{ij} \neq \frac{f_{i.} \cdot f_{.j}}{N}$$

Vejamos:

$$f_{11} = 5 \neq \frac{f_{1.} \cdot f_{.j}}{N} = \frac{35.19}{100} = 6,65$$

PROPRIEDADE:

Se X e Y são independentes então $\mu_{xy} = \mu_x \cdot \mu_y$, onde

$$\mu_{xy} = \frac{1}{N} \sum_i \sum_j x_i \; y_j \; f_{ij} \quad ; \quad \mu_x = \frac{1}{N} \Sigma \, f_{i.} x_i \quad ; \quad \mu_y = \frac{1}{N} \Sigma \, f_{j.} y_j$$

EXEMPLOS:

a) Com os dados da Tabela 12:

$$\mu_{xy} = \frac{1}{100} (0.0.10+0.1\ 20.+0.2.20+1.0.4+1.1.8+1.2.8+2.0.6+$$

$$2.1.12+2.2.12) = \frac{8+16+24+48}{100} = \frac{96}{100} = 9,6$$

$$\mu_x = \frac{1}{100}(0.50+1.20+2.30) = \frac{80}{100} = 0,8$$

$$\mu_y = \frac{1}{100}(0.20+1.40.2.40) = \frac{120}{100} = 1,20$$

Portanto: $\mu_{xy} = 9,6 = 0,8 \ . \ 1,20 = \mu_x \ . \ \mu_y$.

b) Para os dados da Tabela 13, verificamos que:
$\mu_{xy} = 1,22$; $\mu_x = 0,95$ e $\mu_y = 1,32$. Logo, $\mu_{xy} = \neq \mu_x . \mu_y$.

IMPORTANTE: A recíproca da propriedade acima não é verdadeira. Isto é, se $\mu_{xy} = \mu_x \ . \ \mu_y$ não implica que X e Y sejam independentes.

EXEMPLO: Tabela 13:

X \ Y	0	1	2	Σ
0	10	15	10	35
1	15	0	15	30
2	10	15	10	35
Σ	35	30	35	100

Aqui $\mu_x = 1$; $\mu_x = 1$ e $\mu_y = 1$
Portanto: $\mu_{xy} = \mu_x \ . \ \mu_y$. Mas facilmente verificamos que X e Y não são independentes, pois existe pelo menos um f_{ij} tal que $f_{ij} \neq (f_{i.} . f_{.j})/N$. Vejamos: $f_{11} = 10 \neq (f_{1.} \, f_{.1})/100$

3. COVARIÂNCIA ENTRE X E Y

Notação: Cov(X,Y) ou σ_{xy}

A covariância mede a variação conjunta de X e Y. Se a covariância é positiva, o relacionamento é do tipo direto, isto é, para pequenos valores de X correspondem pequenos valores de Y. Analogamente, se ela é negativa, o relacionamento é do tipo inverso, ou seja, para pequenos valores de uma variável correspondem grandes valores da outra e vice-versa. Ela será nula ou próxima de zero para fracos relacionamentos lineares.

Dada uma distribuição bidimensional:

$x_i \backslash y_j$	y_1	y_2	y_3	$\ldots$	y_m	Σ
x_1	f_{11}	f_{12}	f_{13}	$\ldots$	f_{1m}	$f_{1.}$
x_2	f_{21}	f_{22}	f_{23}	$\ldots$	f_{2m}	$f_{2.}$
x_3	f_{31}	f_{32}	f_{33}	$\ldots$	f_{3m}	$f_{3.}$
$\vdots$	$\vdots$					
x_n	f_{n1}	f_{n2}	f_{n3}	$\ldots$	f_{nm}	$f_{n.}$
Σ	$f_{.1}$	$f_{.2}$	$f_{.3}$	$\ldots$	$f_{.m}$	N

definimos covariância entre X e Y:

$$\text{Cov}\ (X,Y) = \sigma_{xy} = \frac{1}{N} \sum_i \sum_j (x_i - \mu_x)\ (y_j - \mu_y)\ f_{ij}$$

Demonstra-se que:

$$\text{Cov}(X,Y) = \sigma_{XY} = \mu_{xy} - \mu_x \cdot \mu_y$$

onde

$$\mu_{xy} = \frac{1}{N} \sum_i \sum_j x_i\ y_j\ f_{ij} \quad ; \quad \mu_x = \frac{1}{N} \Sigma\ f_{i.}\, x_i \quad ; \quad \mu_y = \frac{1}{N} \Sigma\ f_{j.}\, y_j$$

Vejamos a demonstração para dados não agrupados:

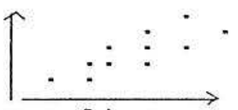
$$\begin{aligned}
\text{Cov}\ (X,Y) &= \frac{1}{N} \Sigma\ (x - \mu_x)\ (y - \mu_y) = \\
&= \frac{1}{N} \Sigma\ (x\ y - y\ \mu_x - x\ \mu_y + \mu_x\ \mu_y) = \\
&= \frac{1}{N} (\ \Sigma\ x\ y - \mu_x\ \Sigma\ y - \mu_y\ \Sigma\ x + \Sigma\ \mu_x\ \mu_y\ ) = \\
&= \frac{\Sigma\ xy}{N} - \frac{\Sigma\ y}{N} \mu_x - \mu_y \frac{\Sigma\ x}{N} + \frac{1}{N}\ N\ \mu_x\ \mu_y = \\
&= \frac{\Sigma\ xy}{N} - \mu_y\ \mu_x - \mu_y\ \mu_x + \mu_x\ \mu_y = \\
&= \mu_{xy} - \mu_x\ \mu_y
\end{aligned}$$

OBSERVAÇÃO 1: Se X e Y são linearmente independentes:

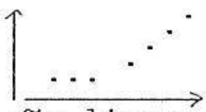
$\mu_{XY} = \mu_X \, \mu_Y$ e, portanto, Cov(X,Y)=0.

Se Cov(X,Y)=0 então X e Y são linearmente independentes, mas isso não significa que sejam sempre independentes.

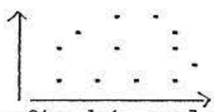
OBSERVAÇÃO 2: Para melhor visualizar a relação, costuma-se construir um gráfico chamado diagrama de dispersão:



linear



não linear



não há relação

EXEMPLO 1: Resultados de uma pesquisa com 10 famílias de determinada região:

poupança (y)	renda (x)	y^2	x^2	xy
4	10	16	100	40
7	15	49	225	105
5	12	25	144	60
20	70	400	4900	1400
20	80	400	6400	1600
30	100	900	10000	3000
8	20	64	400	160
8	30	64	900	240
3	10	9	100	30
15	60	225	3600	900
Σ 120	407	2152	26769	7535

Cov(x,y) = 1/10(7535) - [1/10(407)].[1/10(120]= 265,1

EXEMPLO 2: Distribuição da renda e número de filhos em 40 famílias pesquisadas:

nº filhos / renda (S.M)	0	1	2	Σ
1	1	4	5	10
2	3	5	7	15
5	4	5	6	15
Σ	8	14	18	40

Calcular e interpretar a Cov(X,Y).

4. CORRELAÇÃO - NOÇÕES DE AJUSTAMENTO

Na prática, constata-se frequentemente a existência de uma relação entre duas ou mais variáveis e se deseja expressar tal relação sob forma matemática, estabelecendo-se uma equação. Por exemplo, o peso pode estar relacionado com a idade das pessoas; as vendas de uma empresa podem se relacionar com os gastos promocionais; a demanda de um determinado produto pode estar relacionado como o seu preço.

O problema de determinar equações de curvas que se ajustem a determinados dados é chamado **ajustamento.** Um dos principais objetivos do ajustamento é estimar uma das variáveis(dependente:Y) em função da outra (independente:X). Isto se chama **regressão.**

Podemos também medir o grau de associação ou interdependência entre duas ou mais variáveis através do estudo da **correlação.**

4.1. CORRELAÇÃO

O estudo da correlação tem por objetivo medir e avaliar o grau de relação existente entre duas variáveis. Por exemplo: óbitos por coqueluche e médias semanais de temperatura. A coqueluche é uma doença infecto-contagiosa de alta incidência e cuja letalidade depende de complicações bronco-pneumônicas que ocorrem com maior frequência no inverno, isto é, os óbitos por coqueluche aumentam à medida em que baixa a temperatura.

Outras situações de interesse:

- incidência de gastroenterite abaixo de 2 anos e condições climáticas (aumenta no verão);
- médias de temperaturas e obituário por doenças no aparelho respiratório;
- cólera e condições climáticas;
- pulsação e temperatura de indivíduos normais;

- peso de recém-nascidos e condições sociais dos pais;
- mortalidade infantil e % de pessoas que vivem no mesmo domicílio;
- volume de álcool ingerido diàriamente e duração de vida;
- incidência de uma doença transmissível e imunização da população;
- pulso e temperatura.

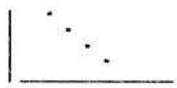
A correlação linear procura medir a relação entre as variáveis X e Y através da disposição dos pontos (X,Y) em torno de uma reta.

COEFICIENTE DE CORRELAÇÃO LINEAR "MOMENTO-PRODUTO" DE BRAVAIS-PEARSON

Notação: ρ_{xy} (população) $\qquad r_{xy}$ (amostra)

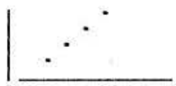
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \ \sigma_y} \quad , \qquad \rho_{xy} \in [-1;1]$$

$\rho_{xy} = -1 \Rightarrow \sigma_{xy} < 0 \Rightarrow \mu_x \ \mu_y > \mu_{xy} \Rightarrow$ há correlação linear inversa perfeita



$\rho_{xy} = 0 \Rightarrow \sigma_{xy} = 0 \Rightarrow$ não há correlação linear

$\rho_{xy} = 1 \Rightarrow \sigma_{xy} > 0 \Rightarrow$ correlação linear direta perfeita



Para os valores intermediários, podemos considerar:

ρ_{xy}	correlação
0	nula
0 —| |0,3|	fraca
|0,3|—| |0,6|	média
|0,6|—| |0,9|	forte
|0,9|—| |0,99|	fortíssima
|1|	perfeita

OBSERVAÇÃO 1: ρ segue o sinal do coeficiente angular da reta de regressão.

OBSERVAÇÃO 2: ρ não informa qual variável influi e qual é influenciada, pois trocando X por Y o valor de ρ não se altera.

OBSERVAÇÃO 3: O coeficiente de correlação é uma quantidade adimensional.

OBSERVAÇÃO4: Se X e Y são independentes então $\rho=0$. Isto decorre do fato de que se X e Y são independentes então $\mu_{xy}=\mu_x\mu_y \Rightarrow$ $Cov(X,Y)= \mu_{xy} - \mu_x.\mu_y = 0$.

COMENTÁRIO: A recíproca da Observação 4 em geral não é verdadeira. Isto é, podemos ter $\rho=0$ e, no entanto, X e Y não precisam ser independentes. Se $\rho=0$, diremos que X e Y são não correlacionadas. Portanto, ser não correlacionados e ser independente, em geral, não são equivalentes. A Tabela 13 serve para ilustrar essa situação: temos que $\mu_{xy}=\mu_x\mu_y$ e, portanto, $\rho=0$, mas X e Y não são independentes.

FÓRMULA PRÁTICA PARA CÁLCULO DE ρ:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \ \sigma_y} = \frac{\mu_{xy} - \mu_x \ \mu_y}{\sigma_x \ \sigma_y} =$$

$$= \frac{\frac{1}{N} \sum_i \sum_j x_i \ y_j \ f_{ij} - \frac{1}{N} \sum f_{i.} x_i \ . \ \frac{1}{N} \sum f_{j.} y_j}{\sqrt{\left[\frac{1}{N} \sum f_{i.} x_i^2 - \mu_x^2 \right] \left[\frac{1}{N} \sum f_{j.} y_j^2 - \mu_y^2 \right]}} =$$

$$= \frac{\frac{1}{N} \sum_i \sum_j x_i \ y_j \ f_{ij} - \frac{1}{N} \sum f_{i.} x_i \ . \ \frac{1}{N} \sum f_{j.} y_j}{\sqrt{\left[\frac{1}{N} \sum f_{i.} x_i^2 - \left(\frac{1}{N} \sum f_{i.} x_i \right)^2 \right] \left[\frac{1}{N} \sum f_{j.} y_j^2 - \left(\frac{1}{N} \sum f_{j.} y_j \right)^2 \right]}}$$

Multiplicando por N^2:

$$= \frac{\frac{N^2}{N} \sum_i \sum_j x_i \; y_j \; f_{ij} - \frac{N}{N} \sum f_{i.} x_i \cdot \frac{N}{N} \sum f_{j.} y_j}{\sqrt{\left[\frac{N}{N}^2 \sum f_{i.} x_i^2 - \left(\frac{N}{N} \sum f_{i.} x_i \right)^2 \right] \left[\frac{N}{N}^2 \sum f_{j.} y_j^2 - \left(\frac{N}{N} \sum f_{j.} y_j \right)^2 \right]}}$$

(No denominador N^2 entra N^4 na raiz: N^2 para x e N^2 para y).

Finalmente:

$$\rho_{xy} = \frac{N \sum_i \sum_j x_i \; y_j \; f_{ij} - \left[\sum f_{i.} x_i\right] \left[\sum f_{j.} y_j\right]}{\sqrt{\left[N \sum f_{i.} x_i^2 - \left[\sum f_{i.} x_i\right]^2\right] \left[N \sum f_{j.} y_j^2 - \left[\sum f_{j.} y_j\right]^2\right]}}$$

EXEMPLO: Calcule o coeficiente de correlação linear entre Renda familiar e Poupança das 10 famílias do Exemplo 1.

Temos que $\Sigma x = 407$, $\Sigma y = 120$, $\Sigma x^2 = 26769$, $\Sigma y^2 = 2152$ e $\Sigma xy = 7535$. Então:

$$\rho_{xy} = \frac{10\ (7535) - (407)(120)}{\sqrt{[10.26769-(407)^2][10.2152-(120)^2]}} = 0{,}9835$$

Existe uma forte correlação linear entre renda e poupança familiar. O sinal do coeficiente mostra que as duas variáveis variam no mesmo sentido.

EXEMPLO: Calcule o coeficiente de correlação linear entre a Renda e o Número de Filhos das 40 famílias pesquisadas no Exemplo 2.

Temos que $N = 40$, $\Sigma\ fx = 50$, $\Sigma\ fx^2 = 86$, $\Sigma\ fy = 115$, $\Sigma\ fy^2 = 445$ e $\Sigma\ fxy = 137$, então $\rho_{xy} = -0{,}13$, o que indica que existe uma fraca correlação linear inversa entre as variáveis. Quanto maior a renda das famúlias, menor o número de filhos.

EXEMPLO: Considere as três situações hipotéticas abaixo onde X é a variação do pulso e Y é a temperatura de pacientes. Trace o diagrama de dispersão para cada grupo de 6 pacientes e calcule ρ_{xy} para cada um.

(a) Pulso	Temp	(b) Pulso	Temp	(c) Pulso	Temp
60	36	60	41	70	40
70	37	70	40	60	41
80	38	80	39	100	38
90	39	90	38	90	37
100	40	100	37	80	36
110	41	110	36	110	39
$\rho_{xy} = 1$		$\rho_{xy} = -1$		$\rho_{xy} = -0{,}37$	

4.2. AJUSTAMENTO - REGRESSÃO

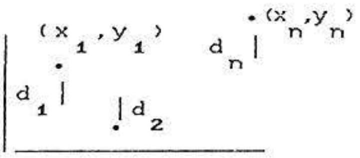
Na prática, o próprio diagrama de dispersão sugere o tipo de curva a ser adotado: reta, parábola, variáveis transformadas por log, etc.

Para um dado X não observamos necessàriamente um Y, então: $Y = f(X) + \varepsilon$ onde ε é a variável que irá captar toda a influência sobre Y não devidas a X. Assim, se a função ajustante é linear, temos que $Y = \alpha + \beta X$ é a regressão de Y sobre X onde Y é variável aleatória e X é a variável sem erro. Nosso problema está em encontrar os valores para α e β

4.2.1. MÉTODO DOS MÍNIMOS QUADRADOS

De um modo geral, pode-se ajustar mais de uma curva a determinado conjunto de dados. A fim de evitar critérios individuais na escolha de retas, parábolas, etc, é nwcessário se chegar a um acordo quanto ao que se deve entender por "melhor reta".

Assim:



Para determinado valor de x, haverá uma diferença entre o valor y e o correspondente valor ajustado pela curva. Se a quantidade $d_1^2 + d_2^2 + \ldots + d_n^2$ é pequena, teremos um bom ajustamento. Então escolhemos a curva que goza da propriedade $\sum d_i^2$ = mínimo , pois é a melhor curva ajustadora.

Vejamos no caso da reta.

Os valores de y na reta de mínimos quadrados, correspondentes a x_1 , x_2 , ...,x_n são $\alpha + \beta\, x_1$, $\alpha + \beta\, x_2$, ..., $\alpha + \beta\, x_n$.

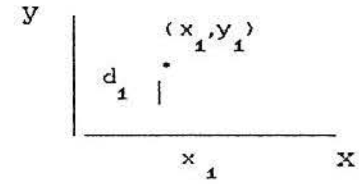
Os correspondentes desvios verticais são:

$d_1 = y_{est} - y_1 = \alpha + \beta\, x_1 - y_1$

$d_2 = y_{est} - y_2 = \alpha + \beta\, x_2 - y_2$

$\vdots$

$d_n = y_{est} - y_n = \alpha + \beta\, x_n - y_n$



Então: $\sum_{i=1}^{n} d_i^2 = \sum (\alpha + \beta x_i - y_i)^2 = f(\alpha,\beta)$ que é função de α e β.

Para que esta função seja mínima: $\dfrac{\partial f}{\partial \alpha} = 0$ e $\dfrac{\partial f}{\partial \beta} = 0$.

Isto é:

$$\frac{\partial f}{\partial \alpha} = \sum \frac{\partial}{\partial \alpha} (\alpha + \beta x - y)^2 = \sum 2 (\alpha + \beta x - y)$$

$$\frac{\partial f}{\partial \beta} = \sum \frac{\partial}{\partial \beta} (\alpha + \beta x - y)^2 = \sum 2\, x (\alpha + \beta x - y)$$

Obtemos então:

$$\sum 2 (\alpha + \beta x - y) = 0 \Leftrightarrow \sum (\alpha + \beta x - y) = 0$$

$$\Sigma\, 2\, x\, (\, \alpha + \beta x - y\,) = 0 \Leftrightarrow \Sigma\, x\, (\, \alpha + \beta x - y\,) = 0$$

Daí :

$$\Sigma\, (\alpha+\beta x-y) = 0 \Leftrightarrow \Sigma\, \alpha + \beta\, \Sigma\, x - \Sigma\, y = 0 \quad \Leftrightarrow \Sigma\, y = N\, \alpha + \beta\, \Sigma x$$

$$\Sigma x(\alpha+\beta x-y)=0 \Leftrightarrow \Sigma \alpha x + \beta \Sigma x^2 - \Sigma xy = 0 \quad \Leftrightarrow \Sigma xy = \alpha\, \Sigma x + \beta\, \Sigma x^2$$

Isto é:
$$\begin{cases} \Sigma\, y = N\, \alpha + \beta\, \Sigma\, x \\ \Sigma\, x\, y = \alpha\, \Sigma\, x + \beta\, \Sigma\, x^2 \end{cases}$$

que é conhecido como sistema de equações normais para a reta de mínimos quadrados. Observemos que de $y = \alpha + \beta\, x$ obtém-se a $1^{\underline{a}}$ equação aplicando somatório e a $2^{\underline{a}}$ multiplicando por x e somando.

Resolvendo o sistema:

$$\alpha = \frac{(\, \Sigma\, y\,)(\, \Sigma\, x^2\,) - (\, \Sigma\, x\,)(\, \Sigma\, x\, y\,)}{N\, \Sigma\, x^2 - (\, \Sigma\, x\,)^2}$$

$$\beta = \frac{N\, \Sigma\, x\, y - (\, \Sigma\, x\,)\, (\, \Sigma\, y\,)}{N\, \Sigma\, x^2 - (\, \Sigma\, x\,)^2}$$

Calculado β, podemos obter α por: $\alpha = \overline{Y} - \beta\, \overline{X}$.

EXEMPLO: Obter a reta de regressão de y sobre X para a Poupança e a Renda das 10 famílias do Exemplo 1.

Temos que $\Sigma x = 407$, $\Sigma y=120$, $\Sigma x =26769$, $\Sigma y =2152$ e $\Sigma xy = 7535$. Então:

$$\beta = \frac{10(7535) - (407)(120)}{10(26769)-(407)^2} = 0{,}2598 \qquad \Rightarrow \; Y = 1{,}4261 + 0{,}2598\, X$$

$$\alpha = 12 - (0{,}2598)(40{,}7) = 1{,}4261$$

4.2.2. PODER EXPLICATIVO DO MODELO OU COEFICIENTE DE DETERMINAÇÃO

Notação: ρ^2 (pop) r^2 (am)

O poder explicativo do modelo, também chamado coeficiente de determinação, tem por objetivo avaliar a "qualidade" do ajuste. Seu valor fornece a proporção da variação total da variável Y explicada pela variável X através da função ajustada.

$$\rho^2 = \frac{\beta^2 \sigma_x^2}{\sigma_y^2} \qquad 0 \le \rho^2 \le 1$$

Quando $\rho^2 = 0$, a variação explicada de Y é zero, ou seja a reta ajustada é paralela ao eixo da variável X. Se $\rho^2 = 1$, a reta ajustada explica toda a variação de Y. Assim, quanto mais próximo de 1 estiver ρ^2, melhor a qualidade do ajuste da função aos pontos do diagrama de dispersão.

Se o poder explicativo for 98%, significa que 98% das variações de Y são explicadas por X através da função escolhida para relacionar as duas variáveis e 2% são atribuidas a causas aleatórias.

4.2.3. ERRO PADRÃO

$$\varepsilon_{XY} = \sqrt{\frac{\Sigma (y_i - y_{est})^2}{N}}$$

Para o caso da reta:

$$\varepsilon_{XY} = \sqrt{\frac{\Sigma y_i^2 - \alpha \Sigma y_i - \beta \Sigma x_i \; y_i}{N}}$$

SUGESTÃO: Complementar o estudo com Regressão linear por transformação e Regressão polinomial (ajuste por parábola)- "Estatística Básica" Toledo.

5. EXERCÍCIOS.

5.1. Usando os dados do quadro abaixo, onde aparecem telespectadores e não-telespectadores, classificados quanto ao seu grau de realização, calcular: a) a porcentagem dos que não são telespectadores, porém grandes realizadores; b) a porcentagem dos telespectadores que são grandes realizadores; c) a proporção de não-telespectadores que são grandes realizadores; d) a proporção de telespectadores que são grandes realizadores.

Realização versus Assistência a Programas de Televisão

Realização	Situação Não-Telespectadores	 Telespectadores	Σ
Alta	93	46	
Baixa	90	127	
Σ	183	173	

5.2. O quadro seguinte representa a estrutura familiar de crianças negras e brancas para uma dada região:

Estrutura Familiar de Crianças Negras e Crianças Brancas

Estrutura Familiar	Raça da criança Negra	 Branca	Σ
Um genitor	53	59	
Dois genitores	130	167	
Σ	183	226	

Calcular: a) a porcentagem de crianças negras que possuem dois genitores; b) a porcentagem de crianças brancas que possuem dois genitores; c) comparar os dois resultados anteriores; d) a proporção de crianças negras que tem um genitor e) a proporção de crianças brancas que tem um genitor; f)comparar d) e e).

5.3. Distribuição dos pais segundo o Método de Educação dos Filhos e sua Orientação Política

Método de Educação	Orientação Política		
	Conservadores	Moderados	Liberais
Permissivo	7	9	14
Moderado	10	10	8
Autoritário	15	11	5
Total	32	30	27

Construa: a) as distribuições marginais para cada variável;
b) as distribuições condicionais de X/Y_j e Y/X_i;

c) Pela simples inspeção das tabelas podes "sentir" algum tipo de associcção entre Orientação Política e Método de Educação ?

5.4. Para testar se determinada droga era capaz de inibir a absorção de álcool pelo organismo humano, realizou-se um experimento com a participação de 30 voluntários (homens saudáveis,idades entre 25 e 28 anos). Uma hora após a ingestão da droga "milagrosa", todos os voluntários tomaram 2 doses de uísque. Uma hora mais tarde, colheu-se uma amostra do sangue de cada sujeito. Os resultados foram:

Droga "milagrosa"	Teste revelou	
	presença de álcool	ausência de álcool
tomaram	4	16
não tomaram	8	2
Total	12	18

Analise percentualmente esses resultados e conclua quanto à eficácia ou não da droga milagrosa.

5.5 Dois grupos de estudantes fizeram exames finais de Estatística. somente um grupo recebeu preparação formal para o exame; o outro leu o texto recomendado, mas nunca compareceu às aulas.Enquanto que 22 dos 30 membros do primeiro grupo (os "frequentadores") passaram no exame, apenas 10 dos 28 do segundo grupo (os "ausentes") lograram aprovação. Organize a distribuição bidimensional e as distribuições condicionais de frequência relativa. Compare os resultados de aprovação dos dois grupos e conclua.

5.6. Em certo hospital, 20 crianças epilépticas foram submetidas a exames eletroencefalográficos. O objetivo da pesquisa era determinar se havia relação de dependência entre região cerebral (anterior/posterior) do foco epiléptico e hemisffério cerebral (direito/esquerdo). Eis os dados:

Região	Hemisfério Direito	 Esquerdo
Anterior	3	8
Posterior	7	2

O que podes concluir ?

5.7. Os seguintes pares de números são os escores obtidos por 20 vendedores num teste destinado a medir a aptidão para vendas e os escores destinados a medir produtividade de venda num certo período de tempo:

Aptidão	Produtividade		
41	32	38	29
35	20	38	33
34	35	46	36
40	24	36	23
33	27	32	22
42	28	43	38
37	31	42	26
42	33	30	20
30	26	41	30
43	41	41	30

a) Trace o diagrama de dispersão e observe se a realção pode ser tratada como linear.

b) Calcule o coeficiente de correlação e interprete.

c) Se a relação puder ser tratada como linear, ajuste uma reta aos dados.

5.8. Calcule o coeficiente de correlação para os escores do teste de inteligência e notas dados a seguir. Trace o diagrama de dispersão.

T.I.	Média	T.I.	Média
295	3,4	230	3,6
152	1,6	195	1,0
214	1,2	174	2,8
171	1,0	236	2,8
131	2,0	198	1,8
178	1,6	217	2,0
225	2,0	143	1,2
141	1,4	135	2,4
116	1,0	146	2,2
173	3,6	127	2,4

5.9. Uma empresa através do Departamento de Finanças informa o total de vendas e as despesas com propaganda, supondo a existência de uma relação linear entre as variáveis. Obtenha a função que relaciona o total de vendas com as despesas com propaganda.

milhões vendas (Y)	120	190	240	140	180	280	150	115	215	220	320
milhões propagandas (X)	2,5	6,5	11	4,0	8,5	14	6,0	5,0	10	13,5	16

5.10. A energia geotérmica pode ser muito importante, no futuro. Como o total de energia contida em uma libra de água é função de sua temperatura, pode-se inquirir se a água proveniente de veios fundos contém ou não mais energia por libra. Os dados a seguir foram extraídos de um artido sobre sistemas geotérmicos, escrito por A. J. Ellis:

Localização do veio de água	Média Máxima de profundidade	Temperatura Média Máxima
El Tateo, Chile	650	230
Ahuachapan, El Salvador	1000	230
Namafjale, Islândia	1000	250
Larderello, Itália	600	200
Matsukawa, Japão	1000	220
Cerro Prieto, México	800	300
Warakei, Nova Zelândia	800	230
Kizildere, Turquia	700	190
The Geysers, Estados Unidos	1500	250

a) Trace o diagrama de dispersão.

b) Determine a reta de mínimos quadrados.

c) Calcule o coeficiente de correlação e interprete.

d) Qual é o poder explicativo do modelo ?

5.11. Os dados a seguir se referem a pacientes que sofrem de enfisema. Registrou-se o número de anos que cada paciente fumou (x) e um diagnóstico médico subjetivo sobre o grau em que os pulmões dos pacientes foram afetados (Y):

Paciente	Número de anos que fumou	Grau de afecção dos pulmões (0 a 100)
1	25	55
2	36	60
3	22	50
4	15	30
5	68	75
6	39	70
7	42	70
8	31	55
9	28	30
10	33	35

Trace um diagrama de dispersão e verifique se uma reta é um bom ajuste para esses dados.

5.12. Os motoristas que estão suspeitos de dirigirem alcoolizados deverão submeter-se, em breve, a um aparelho que mede o teor de álcool do ar expirado: o dispositivo V.SA

H Príncipe. Os dados a seguir, de 15 motoristas alcoolizados, fornecem uma comparação deste novo dispositivo com o aparelho até agora usado mundialmente (o "bafômetro").

Bafômetro(Y)	Dispositivo V.S.(X)
0,15	0,154
0,10	0,085
0,09	0,079
0,14	0,144
0,08	0,078
0,11	0,078
0,12	0,097
0,10	0,088
0,09	0,082
0,09	0,072
0,09	0,080
0,09	0,092
0,08	0,067
0,08	0,077
0,06	0,053

a) Será que os dados permitem ajustar uma reta de mínimos quadrados entre os resultados do "bafômetro" e do dispositivo V.S ?

b) Represente os dados num gráfico e trace a reta de mínimos quadrados.

5.13. A tabela a seguir mostra a distribuição de frequências das alturas respectivas, X e Y, de 80 pais e seus filhos mais velhos.

a) Construa um diagrama de dispersão.

b) Determine o coeficiente de correlação linear e conclua.

X \ Y	150 ⊢ 160	160 ⊢ 170	170 ⊢ 180	180 ⊢ 190	Σ
150 ⊢ 160	2	3	5	6	16
160 ⊢ 170	4	5	8	9	26
170 ⊢ 180	2	4	5	7	18
180 ⊢ 190	3	5	5	7	20

V- ANÁLISE DE SÉRIES TEMPORAIS

Em muitas situações, as observações de interesse são obtidas em instantes sucessivos de tempo(a cada hora, durante 24 horas) ou então registradas por algum equipamento de forma contínua (como um traçado eletrocardiográfico).

Uma série temporal é um conjunto de valores observados em períodos de tempo ordenados. É representada graficamente por um gráfico em linha.

São exemplos de séries temporais:

a) os valores diários do preço das ações da Pirelli, na Bolsa de Valores do Rio de Janeiro;

b) os valores mensais de temperatura da cidade de Porto Alegre;

c) as quantidades anuais de chuva na cidade de Fortaleza;

d) os valores mensais do índice do IEPE;

e) as alturas de marés no porto de Rio Grande, obtidas através de um marégrafo;

f) o registro de um eletrocardiograma (ECG) de um indivíduo;

g) o registro do movimento da crosta terrestre, obtido através de um sismógrafo.

Quando se analisa uma série temporal, os principais objetivos são:

i) modelagem do fenômeno sob consideração;

ii) obtenção de conclusões em termos estatísticos;

iii) avaliação da adequação do modelo em termos de previsão;

iv) sumário dos dados: descrever o comportamento da série verificando a existência de tendências, ciclos e variações sazonais, construindo histogramas e diagramas de dispersão;

v) investigação do mecanismo gerador da série temporal: na análise de ondas oceânicas, queremos saber como estas ondas foram criadas;

vi) estabelecimento de causalidade: na série de vendas de um produto e a respectiva série de propagandas, deseja-se saber se há uma relação de causa e efeito entre elas;

vii) classificação: obtido o ECG de um indivíduo, queremos determinar se ele apresenta algum problema cardíaco, isto é, vamos classificá-lo como normal ou não, baseado em algum padrão;

viii) controle: muitas situações em ciências físicas, biológicas e sociais, engenharia, medicina, envolvem o conceito de sistema dinâmico, caracterizado por uma série de entrada, uma série de saída e uma função de transferência. Na produção de um tipo de aço, com uma determinada quantidade de carbono especificada a priori, tem-se que ajustar convenientemente a "entrada" (quantidade de ferro, seu teor em carbono, etc) e o "sistema" (quantidade de oxigênio injetada, temperatura do forno, etc.).

A análise de série temporal é o procedimento pelo qual são identificados e separados os fatores relacionados com o tempo que influenciam os valores observados da série. Após identificar esses fatores, podemos utilizá-los na interpretação e na projeção da série temporal.

A abordagem clássica da análise de séries temporais identifica quatro de tais influências ou componentes:

1. Tendência secular (T) : Movimento geral de longo prazo nos valores (Y) da série temporal em um grande período de tempo.

2. Flutuações cíclicas (C): Movimentos oscilatórios, para cima e para baixo, com referência à tendência secular, que tem a duração de diversos anos.

3.Variações sazonais (S) ou estacionais: Movimentos para cima e para baixo, em relação à tendência secular, que são completados dentro de um ano e que se repetem anualmente. Tais variações são identificadas, tipicamente, com base em dados mensais ou trimestrais.

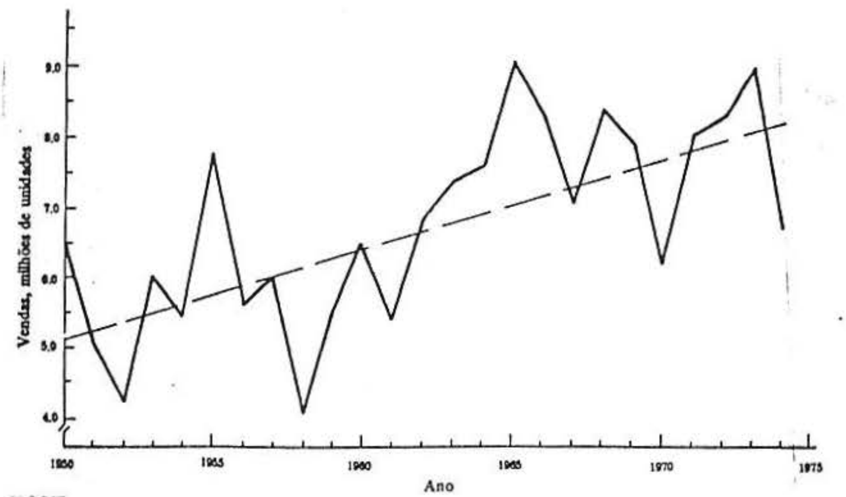
4. Movimentos irregulares (I): Variações erráticas a partir da tendência secular que não podem ser atribuidas a influências cíclicas ou estacionais.

O modelo clássico para a análise de séries temporais é $Y = T \times C \times S \times I$, isto significa que o valor da variável é determinado pelas influências dos quatro componentes e que os componentes tem uma relação multiplicativa.

EXEMPLO 1. Tabela 1. Vendas das fábricas de carros dos Estados Unidos em milhões de unidades, 1950-1974

Ano	Vendas	Ano	Vendas
1950	6,513	1963	7,444
1951	5,090	1964	7,554
1952	4,154	1965	9,101
1953	5,954	1966	8,337
1954	5,352	1967	7,070
1955	7,666	1968	8,407
1956	5,623	1969	7,807
1957	5,953	1970	6,187
1958	4,132	1971	8,122
1959	5,474	1972	8,353
1960	6,530	1973	9,079
1961	5,402	1974	6,721
1962	6,754	Σ	168,779

Graficamente:



Como podemos observar, em 1955 e 1965 ocorreram picos de vendas das fábricas, enquanto que nos anos de 1958 e 1970 as vendas apresentaram a culminância de valores declinantes, constituindo os sulcos.

5.1. ANÁLISE DA TENDÊNCIA

A análise da tendência se refere à direção do movimento de longo prazo na série temporal e é feita utilizando-se os dados anuais. Devem ser usados, no mínimo, 15 anos de dados, de modo que os movimentos cíclicos de diversos anos de duração não sejam tomados como indicativos de uma tendência geral dos valores da série.

O método dos mínimos quadrados é o mais utilizado para identificar a componente tendencial da série temporal. Convém ressaltar que uma linha de tendência não é uma linha de regressão, porque a variável dependente Y não é uma variável aleatória, mas sim um valor historicamente acumulado que apresenta dependência entre períodos adjacentes. Quando o crescimento ou decréscimo de longo prazo parece seguir uma tendência linear, a equação dos valores na linha de tendência, com X representando o anos é $Y_T = \alpha + \beta X$ onde

$$\beta = \frac{N \sum x y - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2} \quad \text{e} \quad \alpha = \overline{Y} - \beta \overline{X}$$

EXEMPLO 2: Ajustar uma reta aos dados do Exemplo 1.

Ano	Vendas (y)	Ano codificado (x)	xy	x^2
1950	6,513	0	0	0
1951	5,090	1	5,090	1
1952	4,154	2	8,308	4
1953	5,954	3	17,862	9
1954	5,352	4	21,408	16
1955	7,666	5	38,330	25
1956	5,623	6	33,738	36
1957	5,953	7	41,671	49
1958	4,132	8	33,056	64
1959	5,474	9	49,266	81
1960	6,530	10	65,300	100
1961	5,402	11	59,422	121
1962	6,754	12	81,048	144
1963	7,444	13	96,772	169
1964	7,554	14	105,756	196
1965	9,101	15	136,515	225
1966	8,337	16	133,392	256
1967	7,070	17	120,190	289
1968	8,407	18	151,326	324
1969	7,807	19	148,333	361
1970	6,187	20	123,740	400
1971	8,122	21	170,562	441
1972	8,353	22	183,766	484
1973	9,079	23	208,817	529
1974	6,721	24	161,304	576
Σ	168,779	300	2194,972	4900

Os anos foram codificados para simplificar os cálculos.

$\overline{X} = 12 \qquad \overline{Y} = 6{,}751 \qquad \beta = 0{,}131 \qquad \alpha = 5{,}179$

Portanto: $Y_T = 5{,}179 + 0{,}131\ X$. Esta equação pode ser utilizada como ponto de partida para projeções. A declividade desta equação linear indica que, durante o período de 25 anos, há uma tendência de longo prazo de um aumento médio anual de aproximadamente 131000 unidades.

No caso de tendência não-linear, frequentemente são utilizadas a curva exponencial (taxa de crescimento constante durante um período de anos) e a curva parabólica.

5.2. MEDIDAS DAS VARIAÇÕES ESTACIONAIS

A influência da componente estacional é identificada através da determinação de um número-índice associado com cada mes (ou trimestre) do ano. A média artimética de todos os 12 números-índices mensais (ou 4 números-índices trimestrais) é 100. A identificação das influências estacionais positivas e negativas é importante para o planejamento de produção e estoques.

Por exemplo, um número-índice de 130 no mes de outubro associado a venda de brinquedos poderia indicar que os valores da série para aquele mes foram 30% mais elevados que os dos outros meses por causa do Dia das Crianças (fator estacional).

O procedimento mais frequentemente utilizado para determinar os números-índices de estacionalidade é o Método da Razão à Média Móvel que consiste basicamente de tres etapas. Primeiro, determina-se o quociente entre cada valor mensal e a média móvel centrada naquele mes. Uma vez que a média móvel baseada em dados mensais (ou trimestrais) para todo o ano "anula", em média, as flutuações estacionais e irregulares, mas não a tendência de longo prazo nem as influências cíclicas, a razão de um valor mensal (ou trimestral) para a média móvel pode ser representada simbolicamente por:

$$\frac{Y}{\text{Média móvel}} = \frac{T \times C \times S \times I}{T \times C} = S \times I$$

O segundo passo do Método da Razão à Média Móvel é cancelar a componente irregular. Isto é feito listando-se as diversas razões aplicáveis ao mesmo mes (ou trimestre) para os diversos anos, eliminando os valores mais alto e mais baixo, e calculando a média das razões restantes. A média resultante é chamada média modificada, por causa da eliminação dos valores extremos.

A etapa final do Método da Razão à Média Móvel consiste em ajustar as razões médias modificadas por um fator de correção, de tal forma que a soma das 12 razões mensais seja 1200 (ou 400 para as quatro razões trimestrais).

EXEMPLO 3:

A Tabela 2 fornece a primeira e principal etapa do Método da Razão à Média Móvel, isto é, o cálculo da razão de cada valor mensal com a média móvel centrada de 12 meses para cada mes.

O primeiro total móvel de 12 meses (6187,4) é o total de unidades vendidas de janeiro até dezembro de 1970. Está centrado entre junho e julho de 1970 por ser um total móvel de um número par de meses.

O total móvel centrado de 2 anos inclui dois períodos sobrepostos de 12 meses. O primeiro total de 2 anos é 12507,9 e inclui o total 6187,4 de janeiro a dezembro de 1970 mais o total 6320,5 de fevereiro de 1970 a janeiro de 1971.

A média móvel de 12 meses é o total móvel centrado de 2 anos dividido por 24.

A média móvel no instante t, para 12 observações, é, portanto, dada por:

$$MM_t^{(12)} = \frac{y_{t-6} + 2\,(y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} +}{24}$$

$$\frac{y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5}) + y_{t+6}}{} = \frac{y_{t-k} + 2\sum y_{t-k+i} + y_{t+k}}{24}$$

Genericamente:

$$MM_t^{(n)} = \frac{y_{t-k} + 2 \sum_{i=1}^{n-1} y_{t-k+i} + y_{t+k}}{2n} \qquad \text{onde} \quad k = n : 2$$

A razão à média móvel é a razão de cada valor de vendas mensais com a média móvel centrada para aquele mes, multiplicada por 100:

$$RMM_t = \frac{y_t}{MM_t} \cdot 100$$

Tabela 2. Vendas mensais de carros, em mil unidades
1970 - 1974

Ano	Mes	Vendas	Total movel 12meses	Total movel centrado 12 meses	Media movel centrada 12 meses	Razao à media movel (%)
1970	jan	545,0				
	fev	528,4				
	mar	594,4				
	abr	627,2				
	mai	684,4				
	jun	758,4				
			6187,4			
	jul	464,3		12507,9	521,16	89,09
			6320,5			
	ago	254,0		12831,6	534,65	47,51
			6511,1			
	set	454,2		13243,7	551,82	82,31
			6732,6			
	out	365,4		13541,6	564,23	64,76
			6809,0			
	nov	341,1		13650,3	568,76	59,97
			6841,3			
	dez	570,6		13685,5	570,23	100,06
			6844,2			
1971	jan	678,1				
	fev	719,0				
	mar	815,9				
	abr	703,6				
	mai	716,7				
	jun	761,3				
	jul	468,9				
	ago	457,6				
	set	712,0				
	out	758,6				
	nov	736,6				
	dez	593,2				

1972	jan	666,0				
	fev	716,1				
	mar	765,2				
	abr	736,9				
	mai	798,0				
	jun	761,6				
	jul	393,6				
	ago	371,0				
	set	808,8				
	out	841,7				
	nov	827,4				
	dez	666,2				
1973	jan	859,8				
	fev	815,5				
	mar	882,8				
	abr	786,6				
	mai	880,1				
	jun	873,3				
	jul	677,5				
	ago	415,7				
	set	666,1				
	out	887,2				
	nov	827,1				
	dez	507,1				
1974	jan	552,1		14849,7	618,74	89,24
			7349,7			
	fev	501,5		14687,5	611,98	81,95
			7343,8			
	mar	557,1		14630,3	609,60	91,39
			7286,5			
	abr	617,4		14448,4	602,02	102,56
			7161,9			
	mai	679,0		13996,3	583,18	116,44
			6834,4			
	jun	618,2		13555,7	564,82	109,46
			6721,3			
	jul	515,2				
	ago	415,8				
	set	608,8				
	out	762,6				
	nov	499,6				
	dez	394,0				

Complete a tabela para os outros anos.

A Tabela a seguir apresenta a segunda e terceira etapas na determinação dos índices de estacionalidade.

Tabela 3. Cálculo dos índices de estacionalidade usando as porcentagens das médias móveis de 12 meses.

Mes	1970	1971	1972	1973	1974	Média modificada por mes	Média modificada x 0,995
Jan		118,86	98,27	112,78	89,24	105,53	105,1
Fev		124,14	106,72	105,08	81,95	105,90	105,4
Mar		136,34	113,97	114,36	91,39		
Abr		112,48	108,54	102,43	102,56		
Mai		108,85	116,30	114,33	116,44		
Jun		112,65	109,90	114,43	109,46		
Jul	89,09	69,38	55,90	91,10			
Ago	47,51	67,73	51,80	57,91			
Set	82,31	105,73	111,51	96,38			
Out	84,77	112,77	114,94	132,31			
Nov	59,98	108,73	112,14	126,25			
Dez	100,07	87,13	89,32	79,72			
					Σ	1205,38	1200

Complete a tabela com os demais dados.

A média modificada por mes é a média das razões à médias móveis para cada mes, depois de serem eliminados os valores mais baixo e mais alto. Assim, para janeiro foram eliminados os valores 89,24 e 118,86, sendo a média dos dois valores restantes 105,53.

As médias modificadas são multiplicadas por um fator de ajustamento, de tal forma que a soma dos índices seja aproximadamente 1200 para os índices estacionais mensais. O fator de ajustamento utilizado com as médias mensais modificadas para obter índices mensais é

$$\text{Fator ajustamento mensal} = \frac{1200}{\text{soma das médias modificadas mensais}}$$

$$\text{OBS:Fator ajustamento trimestral} = \frac{400}{\text{soma das méd. mod trimestr.}}$$

O fator de ajustamento é 1200/1205,38 = 0,9955 que é multiplicado por cada uma das médias modificadas, resultando nos Índices Estacionais Mensais (S_t : última coluna da Tabela 3). Assim, maio é o mes com a maior influência estacional positiva, com a venda de carros sendo 14,8% mais elevada do que no mes típico, em média. Agosto é o mes com a maior componente estacional negativa, com as vendas de carro sendo 54,8% das vendas do mes típico, em média. A última influência está associada com a redução do ritmo de produção das fábricas devido às mudanças anuais de modelos e às escalas de férias, nos Estados Unidos.

5.3. APLICAÇÃO DE AJUSTAMENTOS ESTACIONAIS

Uma aplicação frequente dos Índices Estacionais é o ajustamento dos dados observados da série temporal pela remoção da componente estacional dos dados, o que resulta nos dados estacionalmente ajustados ou dados desestacionalizados. Os ajustamentos estacionais são de interesse para comparar dados de diferentes meses a fim de determinar se um aumento (ou decréscimo) relativo a expectativas teria ou não ocorrido.

Os valores observados mensais (ou trimestrais) da série temporal são ajustados para influências estacionais, dividindo-se cada valor pelo índice mensal (ou trimestral) para aquele mes. O resultado é, então, multiplicado por 100 para manter a posição decimal dos dados originais. O processo de ajustamento de dados para a influência estacional pode ser assim representado:

$$\frac{Y}{S} = \frac{T \times C \times S \times I}{S} = T \times C \times I$$

Embora os valores resultantes da aplicação desta equação estejam na mesma unidade de medida que os dados originais, não representam observações efetivas. São, ao contrário, valores relativos e somente possuem sentido para fins de comparação.

EXEMPLO 4: Desestacionalizar os dados de vendas de carros da Tabela 2, usando os Índices Estacionais Mensais determinados na Tabela 3. Utilizando os resultados da análise, ilustrar o uso de tais valores ajustados estacionalmente.

Para determinar cada valor estacionalmente ajustado, divide-se o valor mensal da Tabela 2 pelo índice de estacionalidade aplicável àquele mes, multiplicando-se o resultado por 100. Assim, o valor ajustado para janeiro de 1970 (518,7) foi obtido dividindo-se 545,0 (da Tabela 2) por 105,1 (da Tabela 3) e multiplicando-se o resultado por 100.

Como utilização dos dados estacionalmente ajustados, note que, em 1974, as vendas de carros pelas fábricas caíram de 618,2 mil unidades para 515,2 mil unidades entre junho e julho (Tabela 2). Contudo, baseados em dados estacionalmente ajustados, as vendas de fato cresceram, como indicado pelos respectivos valores desestacionalizados de 558,0 e 653,2 (Tabela 4). Muito embora tenha ocorrido um decréscimo efetivo, não foi um decréscimo tão grande como seria de se esperar, com base no índice de estacionalidade para aquele mes.

Tabela 4. Dados ajustados estacionalmente para vendas de carros, pelas fábricas dos Estados Unidos, em mil unidades

Mes	1970	1971	1972	1973	1974
Jan	518,7				
Fev	501,1				
Mar	523,0				
Abr					
Mai					
Jun					558,0
Jul					653,2
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

5.4. PROJEÇÕES BASEADAS EM FATORES ESTACIONAIS E EM TENDENCIA

Um ponto de partida para a projeção de longo prazo de valores anuais é fornecido por $Y = \alpha + \beta X$. Contudo, uma consideração particularmente importante em projeções de longo prazo é a componente cíclica da série temporal. Não existe método padrão pelo qual se possa projetar a componente cíclica baseando-se somente em valores históricos da série temporal, mas certos indicadores econômicos são úteis na antecipação dos pontos de reversão do ciclo.

Para projeções de curto prazo, o ponto de partida é o valor projetado de tendência, que é, então ajustado para a componente estacional. Uma vez que a equação para alinha de tendência é geralmente baseada na análise de valores anuais, o primeiro passo necessário é ajustar a equação de forma que ela seja expressa em termos de meses (ou trimestres). Equação de tendência modificada para obter projeções de a.valores mensais:

$$Y_T = \frac{\alpha}{12} + \frac{\beta}{12}\,\frac{X}{12} = \frac{\alpha}{12} + \frac{\beta}{144}\,X$$

b. valores trimestrais:

$$Y_T = \frac{\alpha}{4} + \frac{\beta}{4}\,\frac{X}{4} = \frac{\alpha}{4} + \frac{\beta}{16}\,X$$

Pela transformação para dados mensais, o ponto-base no ano anteriormente codificado como X=0 vai localizar-se agora no meio do ano, ou em 1º de julho. Uma vez que é necessário que este ponto esteja no meio do primeiro mes do ano-base, 15 de janeiro, deve-se reduzir o intercepto $\alpha/12$ da equação modificada por 5,5 vezes a declividade modificada, obtendo-se:

$$Y_T = \frac{\alpha}{12} - 5{,}5\,\frac{\beta}{144} + \frac{\beta}{144}\,X$$

Similarmente, uma equação de tendência modificada para obtenção de valores projetados trimestrais e com X=0 localizado no meio do primeiro trimestre do ano-base é:

$$Y_T = \frac{\alpha}{4} - 1{,}5\,\frac{\beta}{16} + \frac{\beta}{16}\,X$$

EXEMPLO 5: Modificar a equação de tendência,de tal forma que seja expressa em meses, com o mes-base sendo o primeiro mes do ano-base. Utilizar esta equação para determinar os valores mensais de tendência para os anos de 1970 a 1974, em mil unidades.

Para a equação de tendência determinada no Exemplo 2, o ano-base é 1950 (isto é, o ano codificado com X=0). A equação modificada para projetar os valores mensais da tendência, tendo como mes-base janeiro de 1950, é:

$$Y_{T(anual)} = 5,179 + 0,131\ X$$

$$Y_{T(mensal)} = \left[\frac{\alpha}{12} - 5,5\ \frac{\beta}{144} \right] + \frac{\beta}{144}\ X =$$

$$= \left[\frac{5,179}{12} - 5,5\ \frac{0,131}{144} \right] + \frac{0,131}{144}\ X$$

$$= 0,42658\ +\ 0,00091\ X$$

Observemos que os dados anuais da Tabela 1 estão em milhões de unidades e os valores da tendência mensal requeridos são em mil unidades, portanto, a equação fica:

$$Y_{T(mensal)} = 426,58\ +\ 0,91\ X$$

Assim, como janeiro de 1970 é o 240º mes depois do mes-base janeiro de 1950(X=0) o valor de tendência mensal para janeiro de 1970 é:

$$Y_{T(jan\ 1970)} = 426,58 + 0,91\ (240) = 644,98 \cong 645,0$$

Tabela 5. Valores mensais de tendência para as vendas de carros em mil unidades

Mes	1970	1971	1972	1973	1974
Jan	645,0				
Fev	645,9				
Mar	646,8				
Abr					
Mai					
Jun					
Jul					
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

EXEMPLO 6: Determinar os Valores Mensais Previstos das vendas de carro para os anos de 1970 a 1974, aplicando os índices de estacionalidade aos valores de tendência mensais determinados na Tabela 5.

Para janeiro de 1970, o Valor Previsto será de 677,9, obtido multiplicando-se o valor mensal de tendência 645,0 (Tabela 5) pelo Índice de Estacionalidade Mensal 105,1 (Tabela 3) e dividindo-se o resultado por 100. Analogamente para os demais:

$$VP_t = \frac{y_t \cdot S_t}{100}$$

Tabela 6. Valores Mensais Projetados para as vendas de carros com base nas componentes de tendência e estacionalidade, em mil unidades

Mes	1970	1971	1972	1973	1974
Jan	677,9				
Fev	680,8				
Mar	735,4				
Abr					
Mai					
Jun					
Jul					
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

5.5. ANÁLISE DAS VARIAÇÕES CÍCLICAS E IRREGULARES

5.5.1. RELATIVOS DE CICLO

Os valores anuais de uma série temporal representam apenas os efeitos das componentes tendência e ciclo, uma vez que as componentes estacional e irregular são definidas como influências de curto prazo. Portanto, para dados anuais, a componente cíclica pode ser identificada, dividindo-se os valores observados pelo valor de tendência associado, ou seja:

$$\frac{Y}{Y_T} = \frac{T \times C}{T} = C$$

A razão é multiplicada por 100 para que o relativo médio do ciclo seja 100. Um relativo de ciclo 100 indicaria a ausência de qualquer influência cíclica no valor anual da série temporal.

EXEMPLO 7: Determinar os Relativos de Ciclo para os dados da Tabela 1.

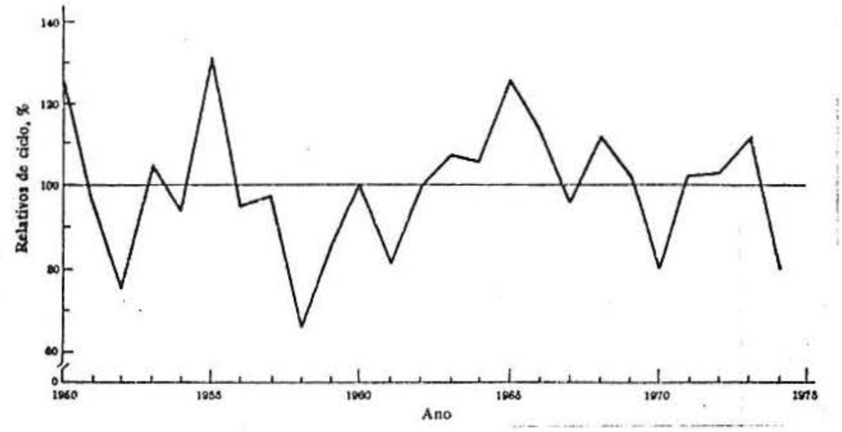
Cada Relativo de Ciclo é determinado multiplicando-se por 100 o valor observado da série temporal (y_t) e dividindo-se o resultado pelo valor da tendência (y_T).

Tabela 7. Relativos de ciclo para as vendas

Ano	Ano cod	Vendas Y	Vendas esper Y_T	Rel. de ciclo 100 Y/Y_T
1950	0	6,513	5,179	125,8
1951	1	5,090	5,310	95,9
1952	2	4,154	5,441	76,3
1953	3	5,954		
1954	4	5,352		
1955	5	7,666		
1956	6	5,623		
1957	7	5,953		
1958	8	4,132		
1959	9	5,474		
1960	10	6,530		
1961	11	5,402		
1962	12	6,754		
1963	13	7,444		
1964	14	7,554		
1965	15	9,101		
1966	16	8,337		
1967	17	7,070		
1968	18	8,407		
1969	19	7,807		
1970	20	6,187		
1971	21	8,122		
1972	22	8,353		
1973	23	9,079		
1974	24	6,721		

Para auxiliar na interpretação dos Relativos de Ciclo, elabora-se, frequentemente um diagrama de ciclo que retrata os relativos de ciclo de acordo com o ano. Pela construção deste diagrama, podem tornar-se mais evidentes os picos e sulcos da componente cíclica das séries temporais.

EXEMPLO 8: Traçar um diagrama de ciclo para as vendas das fábricas de carros de 1950 a 1974, tendo como referência os relativos de ciclo da Tabela 7.



5.5.2. COMPONENTES CÍCLICA E IRREGULAR

Valores mensais ou trimestrais de séries temporais incluem a .influência de todas as quatro componentes das séries temporais: tendência, ciclo, estacionalidade e irregularidade. Pelo emprego do método residual, os efeitos associados com as componentes cíclica e irregular são identificados pela remoção sistemática das influências associadas com as outras componentes da série temporal. O primeiro passo é remover os efeitos da componente tendencial e estacional, dividindo os valores mensais pelos valores projetados com base na componente de tendência e estacionalidade:

$$\frac{Y}{T \times S} = \frac{T \times C \times S \times I}{T \times S} = C \times I$$

Esta razão é multiplicada por 100, de tal forma que possa ser interpretada como uma porcentagem.

EXEMPLO 9:Utilizar o método dos resíduos para identificar as componentes cíclica e irregular combinadas para os dados da série em estudo.

Tabela 8. Valores mensais ajustados para as componentes de tendência e estacionalidade (indicando a influência dos efeitos cíclicos e irregulares)

Mes	1970	1971	1972	1973	1974
Jan	80,4	98,4	95,0	120,7	76,3
Fev	77,6				
Mar	80,8				
Abr	92,1				
Mai	91,9				
Jun	105,4				
Jul					
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

O indice de 80,4 para janeiro de 1970 é obtido dividindo-se as vendas mensais, em mil unidades, de 545,0 (Tabela 2) pela previsão baseada nas componentes tendencial e estacional 677,9 (Tabela 6), multiplicando-se o resultado por 100. Indica que as influências combinadas das componentes cíclica e irregular para aquele mes resultaram em unidades de venda que foram 19,6 % mais baixas do que as vendas previstas com base em fatores de tendência e de estacionalidade.

5.5.3. COMPONENTE CÍCLICA

Para obter a Componente Cíclica, calcula-se uma média móvel dos valores residuais que representam as componentes cíclica e irregular, geralmente, para períodos de 5 meses. Tal média móvel "cancela", em média, a componente irregular de curto prazo, deixando somente a componente cíclica.

A média móvel no mes t para períodos de 5 meses é dada por

$$MM_t^{(5)} = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5} = MM_{t-1} + \frac{y_{t+2} - y_{t-3}}{5}$$

O termo "média móvel" é utilizado porque, a cada período, a observação mais antiga é substituída pela mais recente, calculando-se uma nova média.

EXEMPLO 10: Com referência aos resíduos identificados no Exemplo anterior, identificar a componente cíclica nos valores mensais. Houve picos ou depressões durante o período de 1970 a 1975 ?

Tabela 9. Médias móveis de 5 meses de valores mensais ajustados para componentes de tendência e estacionalidade (indicando a influência percentual dos efeitos cíclicos)

Mes	1970	1971	1972	1973	1974
Jan					
Fev					
Mar	84,6				
Abr	89,6				
Mai	92,1				
Jun	90,3				
Jul					
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

Esta Tabela apresenta as médias móveis de 5 meses para os dados da Tabela 8. Podemos observar que os anos nos quais houve uma influência negativa substancial, associada com a componente cíclica da série temporal foram 1970 e 1974. Os resultados de 1970 coincidem com uma recessão nacional, enquanto os resultados de 1974 coincidem tanto com uma crise de energia como com uma recessão nacional. O ano no qual houve a maior influência positiva associada com a componente cíclica em tais valores mensais foi 1973, que se seguiu à recuperação da recessão de 1969-1970 e que precedeu a crise energética. No ambiente inflacionário de 1973, muitas vendas de carros foram também estimuladas pela antecipação de preços substancialmente mais elevados que seriam cobrados pelos automóveis modelo 1974.

5.5.4. COMPONENTE IRREGULAR

A Componente Irregular de curto prazo pode ser identificada nos dados históricos, dividindo-se os resíduos que contém as componentes cíclica e irregular pela média móvel que representa apenas a componente cíclica:

$$\frac{C \times I}{C} = I$$

EXEMPLO 11: Determinar a componente irregular para as vendas mensais dos carros, referindo-se aos resíduos determinados nos Exemplos 9 e 10.

Tabela 10. Variações percentuais atribuíveis à componente irregular da série temporal

Mes	1970	1971	1972	1973	1974
Jan					
Fev					
Mar	95,5				
Abr	102,8				
Mai	99,8				
Jun	116,7				
Jul					
Ago					
Set					
Out					
Nov					
Dez					

Complete a Tabela.

Os índices são obtidos dividindo-se os índices da Tabela 8, que incluem as influências das componentes cíclica e irregular da série temporal, pelos respectivos índices da Tabela 9, que incluem somente a influência da componente cíclica e multiplicando-se o resultado por 100. Esses índices indicam a variação percentual nos valores da série temporal mensal atribuível à componente irregular. As variações (de 100) que são relativamente grandes podem ser explicadas geralmente pelo conhecimento dos eventos causadores específicos. Por exemplo, uma greve na General Motors durante outubro e novembro de 1970 resultou numa redução substancial nas vendas das fábricas de carros durante esses meses, e o estabelecimento de um acordo nesta disputa trabalhista foi seguido por um momento incomum nas vendas das fábricas em dezembro de 1970 (124,5).

5.6. PREVISÃO DE CICLOS E INDICADORES ECONÔMICOS

A projeção baseada nos componentes tendencial e estacional de uma série temporal é considerada apenas um ponto de partida na previsão econômica. Razões para isso: necessidade de levar em consideração os efeitos prováveis da componente cíclica durante o período de projeção e de identificar os fatores causais específicos que exerceram influência sobre as variáveis das séries temporais.

Para projeção de curto prazo, supõe-se frequentemente que o efeito da componente cíclica é o mesmo incluído em valores recentes da série temporal. Contudo, para períodos longos, ou mesmo para períodos curtos durante épocas de instabilidade econômica, a identificação dos pontos de reversão do ciclo para a economia nacional é bastante importante. É claro que as variações cíclicas associadas com um produto particular podem ou não coincidir com o ciclo geral dos negócios.

VI - ANÁLISE EXPLORATÓRIA

1. INTRODUÇÃO

O trabalho do Estatístico está relacionado com a solução de problemas. Por isso, ele precisa escolher a(s) técnica(s) estatística(s) mais adequada(s) para cada caso. As técnicas estatísticas foram delineadas para serem as melhores possíveis sob rigorosas suposições. Corre sério risco aquele que irresponsavelmente, utiliza uma técnica estatística sem verificar se as suposições exigidas são satisfeitas.

A análise exploratória de dados fornece um conjunto de métodos para sondagem dos dados, antes de adaptá-los a algum modelo probabilístico. Constitui uma fase preliminar de busca de indicação de possíveis modelos a serem utilizados,serve para detectar as suposições feitas sobre o comportamento probabilístico das variáveis são satisfeitas. Ajuda a detectar a presença de valores espúrios (outliers), afastamentos importantes de suposição de normalidade, heteroscedasticidade (variância diferentes), etc, que, em geral, são exigências da maioria das técnicas estatísticas clássicas.

2. RAMO E FOLHAS

Ao se analisar um conjunto de valores numéricos é em geral desejável (ou necessário) dispor desses valores em forma ordenada e além disso de forma que se tenha uma idéia boa sobre o conjunto desses valores. E isso deve em geral preceder qualquer tentativa de sumarizar o comportamento desse conjunto de valores.

Uma maneira simples e eficiente de se conseguir o que é proposto acima pode ser conseguido através da apresentação dos valores num RAMO E FOLHAS.

Considere-se , por exemplo, as populações dos estados brasileiros em 1970, em milhares de habitantes:

(Acre, 218);(Amazonas, 961); (Pará, 2.197);(Maranhão,3.037); (Piauí, 1.735); (Ceará, 4.492); (R. G. do Norte, 1.612);

(Paraíba,2.445); (Pernambuco, 5.253); (Alagoas, 1.606); (Sergipe, 911); (Bahia,7.583); (M. Gerais, 11.645); (E. Santo, 1.618); (R.Janeiro, 9.110); (S. Paulo, 17.959); (Paraná, 6.998); (S. Catarina, 2.930); (R. G. do Sul,6.755) (Goiás, 2.998); (M. Grosso, 1.624).

Vamos quebrar cada valor da seguinte maneira:

	Ramo	Folha	Esquece	Milhares de Hab.
Acre	0	2	18	218
Amazonas	0	9	61	961
Pará	2	1	97	2.197
Maranhão	3	0	37	3.037

Pode-se assim apresentar todo o conjunto de valores, de forma resumida conforme o quadro 1. Nele se tem, de forma resumida, uma boa apresentação do comportamento dos valores de população dos estados brasileiros em centenas de milhares de habitantes.

Vê-se que para "entender" comparativamente os valores de população dos estados foi possível reduzir bastante o número de dígitos original. Não se deve permitir que um amontoado de dígitos atrapalhe ou até impeça o entendimento.

Quadro 1- Ramo e folhas das populações dos estados do Brasil em 1970 em centenas de milhares de habitantes.

0	2 9 9
1	7 6 6 6 6
2	1 4 9 9
3	0
4	4
5	2
6	9 7
7	5
8	
9	1
10	
11	6
12	
13	
14	
15	
16	
17	9

Há uma série de observações que se podem fazer sobre esse conjunto de valores. Certamente não são as únicas.

I) Há um destaque excepcional do valor 179 - correspondente ao estado de São Paulo.

II) Além deste estão bem destacados os valores 116 (Minas Gerais) e 91 (Rio de Janeiro).

III) Os demais valores estão razoavelmente concentrados entre 2 (Acre) e 75 (Bahia).

IV) Um valor (mais ou menos) típico para esse conjunto de valores (excetuados os três destacados) poderia, por exemplo ser 20 - isto é 2 milhões de habitantes. Poder-se-iam fazer também observações de natureza mais formal, como:

V) Há uma acentuada assimetria em direção a valores grandes o que, em conjunto com a impressão geral, permite dizer que:

VI) Não faz sentido supor que tal conjunto de valores pudesse ser, de alguma forma, considerado como proveniente de uma (eventual) população com distribuição normal.

Esta não é a única maneira de representar os ramos e folhas. No caso que está sendo considerado por exemplo, pode-se desejar representar os valores de forma mais compacta, já que estão muito distendidos - pelo menos para valores grandes. Outra forma de apresentação dos mesmos valores é a seguinte, em que cada ramo é constituído por dois valores do caso anterior, como no quadro 2. Nas folhas, deve-se chamar a atenção para que os dois pontos (:) não ocupem o lugar de um dígito, para não distorcer a forma.

<u>Quadro 2</u> - Ramo e folhas (outra versão) das populações dos estados do Brasil em 1970, em centenas e milhares de habitantes

0 : 1	2 9 9:7 6 6 6 6
2 : 3	1 4 9:0
4 : 5	4:2
6 : 7	9 7:5
8 : 9	:1
10 :11	:6
12 :13	
14 :15	
16 :17	:9

É importante observar como a forma de apresentação muda um pouco (às vezes muito) a impressão e altera as observações feitas. Neste caso II) deixa de ser e III) (é claro) se altera. Mas I) ganha ainda mais peso por ter-se mantido.

Uma terceira forma é o "espichamento" da forma original do quadro 1 de forma que cada ramo se desdobra. É apresentado no quadro 3 o correspondente ramo e folhas apenas para os valores até 75 - o que permite de certa forma ver detalhes do conjunto (mais ou menos) compacto de valores menores. Nas folhas vão os valores com dígitos 0, 1, 2, 3, 4 na primeira parte e 5, 6, 7, 8, 9 na segunda.

Quadro 3 - Ramo e folhas (desdobrado) das populações dos estados do Brasil em 1970,em centenas de milhares de habitantes até o valor 75

0	2
0	9 9
1	
1	7 6 6 6 6
2	1 4
2	9
3	0
3	
4	4
4	
5	2
5	
6	
6	9 7
7	
7	5

Para apresentar a quarta forma do ramo e folhas; considere-se outro conjunto de valores:

População das capitais dos estados em milhares de habitantes:

(Rio Branco, 57); (Manaus, 284); (Belém, 565); (S.Luiz,168); (Terezina,181); (Fortaleza,520); (Natal,251); (J.Pessoa,197) (Recife,1.046); (Macei , 243); (Aracajú, 180);(Salvador,998) (B. Horizonte, 1.107); (Vitória, 122); (R. Janeiro,4.252); (S. Paulo, 5.187); (Curitiba, 483); (Florianópolis, 116); (P. Alegre, 870); (Goiânia, 362); (Cuiabá, 88).

Neste caso qualquer tentativa de representar os valores num ramo e folhas das formas anteriores necessitaria de um número excessivo de linhas pois os valores se extendemde 57 a 5.187 (Rio Branco e S.Paulo, respectivamente). Seriam necessárias pelo menos 26 linhas na forma mais compacta e nas últimas 19 linhas só haveria 2 valores.

Uma quarta forma de ramo e folhas é pois apresentada, para situações como esta, no quadro 4, com ramos divididos. A divisão ocorre cada vez que se muda por um fator de dez. O número de situações em que ocorre esse tipo de comportamento é suficientemente elevado a ponto de justificar a existência de um nome (jocoso) para elas de "lei anormal dos grandes números".

Quadro 4 - População das capitais dos estados brasileiros em 1970 em milhares de habitantes

5	7
6	
7	
8	8
9	
1	68,81,97,80,22,16
2	84,51,43
3	62
4	83
5	65,20
6	
7	
8	70
9	98
1	046,107
2	
3	
4	252
5	187

Para enfatizar bem que as técnicas que estão sendo apresentadas não são de maneira alguma fixas poder-se-ia apresentar o mesmo ramo e folhas, bem como no quadro 5, com o ramo zero. Em todas as situações de análise exploratória o fundamental são a atenção, a imaginação e criatividade. Deve-se sempre saber bem o que se está fazendo sem necessariamente prender-se a regras fixas.

<u>Quadro 5</u> - População das capitais - outra vez - com ramo e folhas modificado para economizar quatro linhas

0	57,88
1	68,81,97,80,22,16
2	84,51,43
3	62
4	83
5	65,20
6	
7	
8	70
9	98

1	046,107
2	
3	
4	252
5	187

3.ESQUEMA DE CINCO NÚMEROS

É claro que não é sempre que os valores com que se trabalha estão dispersos assimetricamente para valores grandes. A seguir estão listadas as distâncias de Brasília às capitais dos estados brasileiros em quilômetros, com o ramo e folhas correspondente no quadro 6.

(Rio Branco, 3.033); (Manaus, 3.421); (Belém, 2.100); (S.Luiz, 2.209); (Terezina, 1.751); (Fortaleza, 2.648); (Natal,2.579); (J. Pessoa, 2.408); (Recife, 2.303); (Maceió,2.075); (Aracajú, 1.806); (Salvador, 1.530); (b.Horizonte, 740); (Vitória, 1.278); (R. Janeiro, 1.214); (S.Paulo,1.011); (Curitiba, 1.306); (Florianópolis, 1.608); (P. Alegre, 2.021); (Goiânia, 193); (Cuiabá, 1.127).

<u>Quadro 6</u> - Distância de Brasília às capitais dos estados em dezenas de quilômetros - já ordenadas.

0	19
0	74
1	01,12,21,27,30
1	53,60,75,80
2	02,07,10,20,30,40
2	57,64
3	03,42

Em qualquer caso, agora que se dispõe de uma técnica para apresentar, de forma ordenada, conjuntos de valores, é importante pensar em como resumir de forma útil a informação sobre o comportamento do conjunto que se está considerando. Uma forma usual de fazê-lo é apresentando os valores da média e do desvio padrão. Esses dois valores são, no entanto, não só de pouca utilidade como conduzentes erros de interpretação pois:

I) São afetados de forma exagerada por valores extremos.

Por exemplo, no caso do quadro 1 a média é de 4.461 que vê-se dificilmente pode ser considerado como um valor central das populações dos estados brasileiros.

II) Apenas com essas duas medidas não fica nada do fato de os valores se espalharem assimetricamente para números grandes. E não faz muito sentido usar o valor da assimetria pois será, mais ainda que os anteriores, afetado pelos valores muito extremos.

Uma possível maneira de resumir conjuntos de valores numéricos é proposta a seguir. Serão utilizados:

A MEDIANA - que é um valor que deixa metade abaixo e metade acima dela;

OS QUARTIS - OU JUNTAS - que fazem o mesmo com as duas metades demarcadas pela mediana;

OS EXTREMOS - que são o menor (L_i) e maior (L_s) valores.

A forma de apresentação é então o

ESQUEMA DE CINCO NuMEROS - em que além dos valores mencionados acima aparece de forma explícita o número de valores (n) no conjunto considerado:

	N	
	Md	
Q_1		Q_2
L_i		L_s

Para o caso das populações dos estados brasileiros apresentadas em ramo e folhas no quadro 1 o esquema de cinco números é dado no quadro 7, em duas formas. Em cada caso é opcional conservar ou não os dígitos originais.

Quadro 7 - Duas formas de esquema de cinco números para os valores da população dos estados brasileiros.

21	
*	218
J	1.606
M	2.930
J	6.998
*	17.959

em milhares de habitantes

	21	
M	29	
J	16	69
*	2	179

em centenas de milhares de habitantes

Em qualquer das formas consideradas tem-se uma boa quantidade de informação sobre o comportamento do conjunto como um todo. Por exemplo:

I) A amplitude (que é a diferença de extremos) - 177 - é muito maior do que a diferença de juntas - 53 : mais de três vezes.

II) A mediana - 29 - está muito mais próxima da - junta inferior 16 - do que da junta superior - 69.

Essas duas observações são fortes indicações da assimetria para valores grandes e do grande espalhamento dos valores em si.

4. DESENHO ESQUEMÁTICO

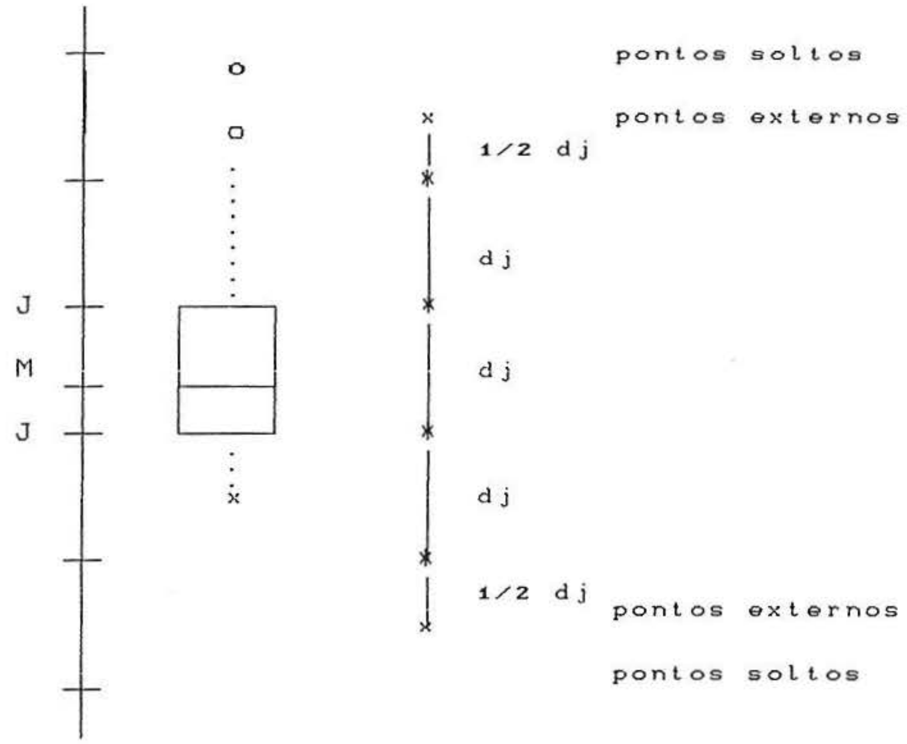
Há um ditado popular que diz: "um desenho vale por mil palavras". Dificilmente há situações em que isso é mais verdadeiro do que em análise exploratória. A seguir é apresentada uma maneira de traduzir em um desenho a informação do esquema de cinco números - e muito mais. Logo a seguir, no quadro 8, é apresentado um desenho esquemático que será descrito gradualmente.

Primeiro, no retângulo central, estão representados as juntas e a mediana. A partir do retângulo, para cima e para baixo, seguem linhas pontilhadas (ou tracejadas) que continuam até:

I) Encontrar um valor extremo, ou

II) Um comprimento dj - diferença de juntas - se o extremo correspondente está a mais de uma dj da junta respectiva.

Quadro 8 - Representação de um desenho esquemático com os pontos e as regiões correspondentes.



A partir daí dá-se destaque especial aos pontos correspondentes a valores do conjunto que se está considerando:

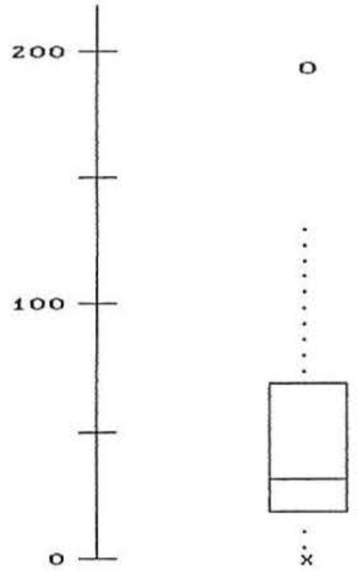
I) Pontos que estão a mais de uma dj da junta correspondente e até uma dj e meia são chamados pontos externos.

II) Pontos que estão a mais de uma dj e meia da respectiva junta são chamados pontos soltos.

Vale observar que se, o conjunto de valores considerado fosse uma amostra de uma população normal, a probabilidade de se observar um ponto externo seria de aproximadamente , 4/100 em cada lado e de observar um ponto solto, 8/1000 (ou menos de um em cem) em cada lado.

O desenho esquemático correspondente aos valores de população dos estados brasileiros, representados no ramo e folhas do quadro 1, está no quadro 9.

Quadro 9 - Desenho esquemático dos valores de população dos estados brasileiros - 1970 - em centenas de milhares de habitantes

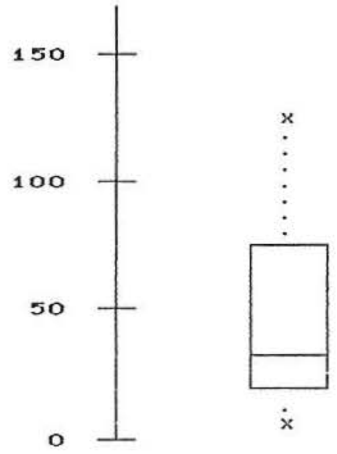


Aqui, mais do que em qualquer outro ponto anterior aparece o real destaque do valor correspondente ao estado de São Paulo. Se temporariamente esse valor for deixado de lado como sendo "excepcional" o desenho esquemático correspondente aos valores restantes é o apresentado no quadro 10. Pode-se realmente considerar que agora já se tem um conjunto de valores "bem comportado" - no sentido de que:

I) a assimetria é menos acentuada

II) não há pontos externos nem soltos.

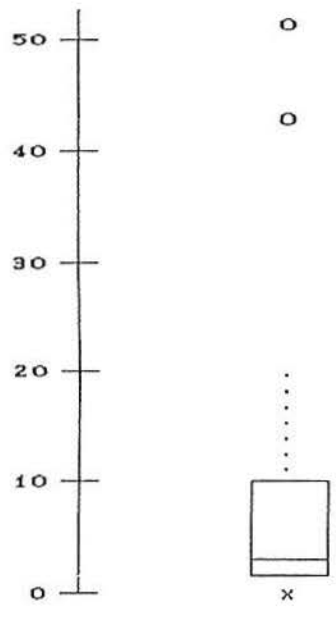
Quadro 10 - Desenho esquemático dos valores de população dos estados brasileiros - 1970 - em centenas de milhares de habitantes - São Paulo excluído.



No ítem 1.4, ainda se irá considerar esse conjunto de valores - aí muito mais sob o prisma de se tentar contornar o problema de assimetria .

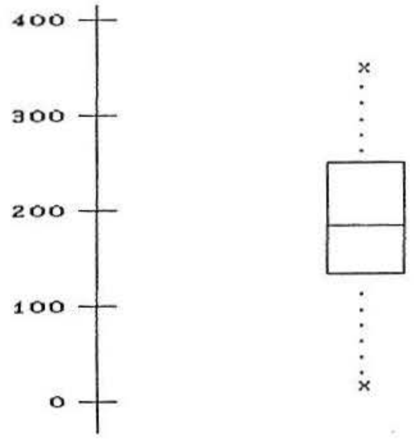
Nos quadros 11 e 12 são apresentados desenhos esquemáticos dos conjuntos apresentados em ramos e folhas nos quadros 4.6.

Quadro 11 - Desenho esquemático correspondente aos valores de polulação das capitais dos estados brasileiros em 1970 em centenas de milhares de habitantes.



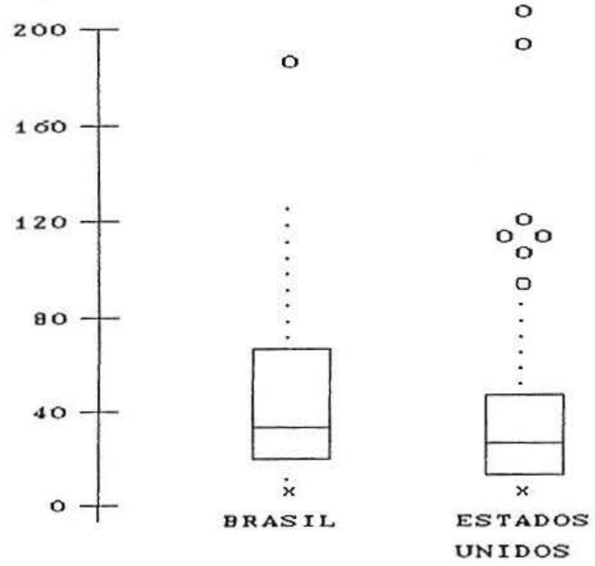
Uma das grandes utilidades do desenho esquemático é também ajudar na comparação de diferentes conjuntos de valores. Apesar de pouco será utilizado para este fim, nestas notas, no quadro 13 estão representados os desenhos esquemáticos correspondentes aos valores de população dos estados brasileiros e dos Estados Unidos em 1970.

Quadro 12 - Desenho esquemático do conjunto de valores das distâncias de Brasília às capitais estaduais em dezenas de quilômetros



Há várias observações que podem ser feitas quanto ao comportamento comparativo dos dois conjuntos de valores considerados. No ítem seguinte serão vistas transformações e lá este exemplo será considerado novamente.

Quadro 13 - Desenhos esquemáticos - comparativos - das populações dos estados do Brasil e dos Estados Unidos em 1970, em centenas de milhares de habitantes.



5. RE-EXPRESSÃO (TRANSFORMAÇÃO)

Um dos melhores exemplos da conveniência - necessidade- de transformar dados é dada por MacNeil:

Uma discussão permanente quando alguém começa a estudar geografia é de se a Austrália é uma ilha ou um continente. Como, trabalhar com as áreas de todas as ilhas é impossível, estão listadas abaixo as áreas de "ilhas" com mais de 10.000 milhas quadradas. Além disso as ilhas pequenas, sejam poucas ou muitas, são irrelevantes para resolver a questão. Ásia e Europa estão listadas separadas. Juntando as áreas de ambas em Eurásia não fará diferença.

AxelH.	16	Baffin	184	Banks	23	Devov	21
Melv.	16	P. of W.	13	South.	16	Victoria	82
Spits	15	Grabret.	84	Irlanda	33	Groel.	840
Terra N.	43	T. d. F.	19	Cuba	43	Hisp.	30
Madag.	227	Formosa	14	Hainan	13	Hohkaiko	30
Kyushu	14	N. Guiné	306	Nova Z.(N)	44	Nova Z (S)	58
Mird.	36	Sakh.	29	Tasm nia	26	Vancouver	12
Celeb	73	Java	49	Moluca	29	N. Br.	15
Timor	13	Elles	82	Nova Z.	32	Isl ndia	40
Ceilão	25	Houshu	89	Luzon	42	Bornéu	280
Sumatra	183	sia	16988	frica	11506	Am. Norte	9390
Am. Sul	6795	Europa	3745	Australia	2968	Antar.	5500

As áreas estão em milhares de milhas quadradas. O quadro 14 dá um ramo e folhas em centenas de milhares de milhas quadradas.

Quadro 14 - Ramo e folhas das áreas das "ilhas" com mais de 10.000 milhas quadradas, em centenas de milhares.

0 : 1	0 0000000000000000000000000000000111111222338:	
2 : 3	: $\boxed{0}$ 7	
4 : 5	: 5	
6 : 7	8:	
8 : 9	: 4	
10 :11	: 5	Austrália
12 :13		
14 :15		
16 :17	: 0	

Este ramo e folhas é pouquíssimo revelador pois os pequenos valores estão muito "embolados". A "distribuição" é extremamente assimétrica. É necessário fazer uma transformação dos dados. Inicialmente será tentada raiz quadrada. O resultado está no quadro 15.

<u>Quadro 15</u> – Área das ilhas com mais de 10.000 milhas quadradas - raízes quadradas das áreas em milhares de milhas quadradas.

```
 0 : 1  | 3 444444444445555555566667777789999:44577
 2 : 3  | 9:
 4 : 5  | :[4]
 6 : 7  | 1 : 4
 8 : 9  | 2 : 7                          Austr lia  [ ]
10 :11  | 7 :
12 :13  | :0
```

A transformação utilizada não foi suficiente para atenuar suficientemente ou eliminar a assimetria. No quadro 16 são apresentados os logaritmos das áreas. É utilizada base 10 e são tomados logaritmos das áreas em milhares de milhas quadráticas.

Agora finalmente as coisas começam a ficar mais claras. Na realidade existem dois aglomerados distintos de valores. E estão a salvo os livros dos australianos. Vista a vantagem - necessidade - de fazer transformações é fundamental entender porque isso acontece.

<u>Quadro 16</u> – Área das "ilhas" com mais de 10.000 milhas quadradas logaritmos em base dez das áreas em milhares de milhas quadradas.

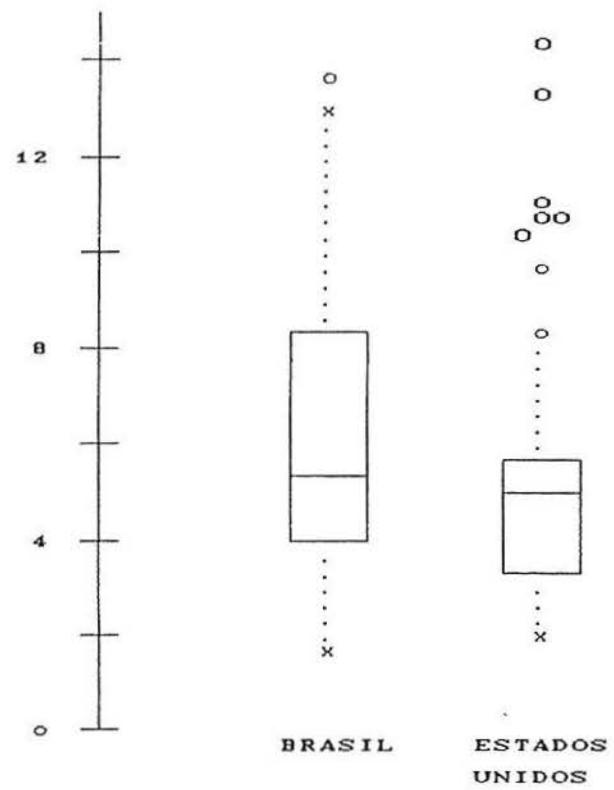
```
1 | 1 11122222223344
1 | 5 55555666666789999
2 | 0 334
2 | 5 59                     Austr lia  [ ]
3 |
3 | [5] 678
4 | 0 12
```

Outra razão da necessidade de fazer transformações está em que, ao compararmos dois (ou mais) conjuntos de valores, se as dispersões são muito diferentes, ou se existe muita assimetria, a comparação é difícil, se não impossível. Em geral, mesmo que não seja imprescindível, uma transformação melhora o entendimento que se tem de um conjunto de valores.

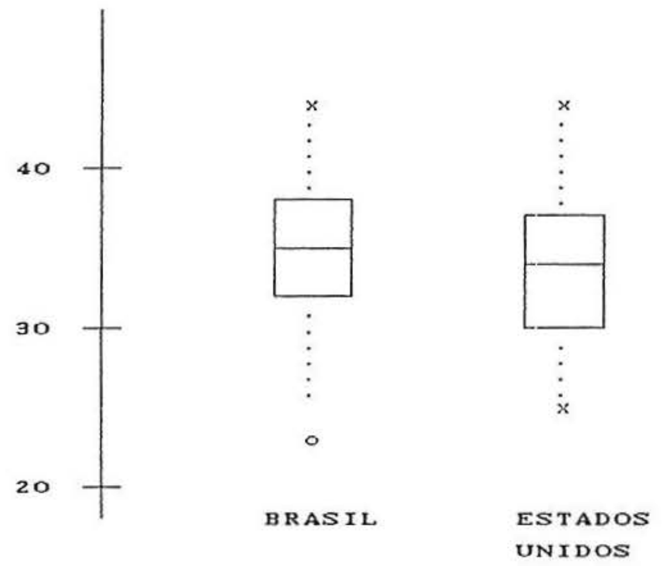
Nos quadros 17 e 18 estão os desenhos esquemáticos correspondentes, respectivamente, as raízes quadradas e aos logaritmos (base 10) dos valores de população (em milhares de habitantes) dos estados brasileiros e dos Estados Unidos. O entendimento que se ganha, para comparar os dois conjuntos, é muito grande ao se fazer as transformações.

Vale à pena, também, chamar a atenção para a importância da simetria. Na realidade onde existem assimetrias muito acentuadas é impossível - pelo menos difícil - apresentar um valor que sirva como indicador de nível para um conjunto de valores. Senão por outras, pelo menos como forma de se atingir alguma assimetria, as transformações são de suma importância.

Quadro 17 - Raízes quadradas das populações dos estados originalmente em milhares de habitantes.

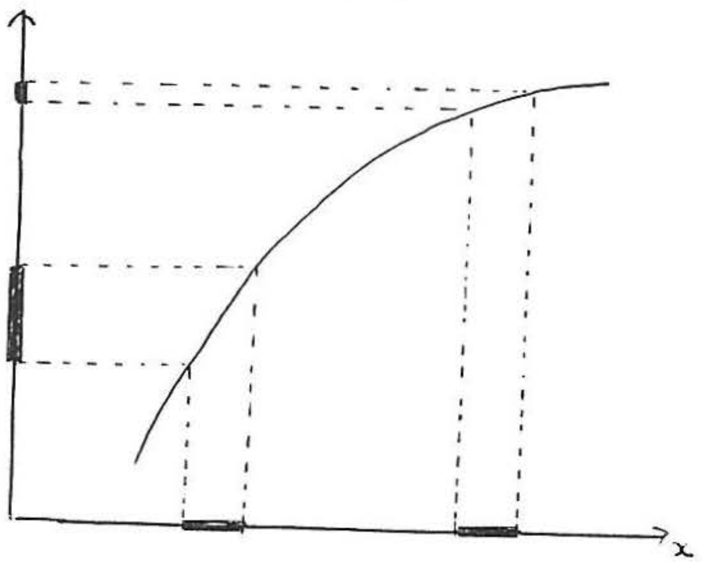


Quadro 18 - Logaritmos das populações dos estados - originalmente em milhares de habitantes.



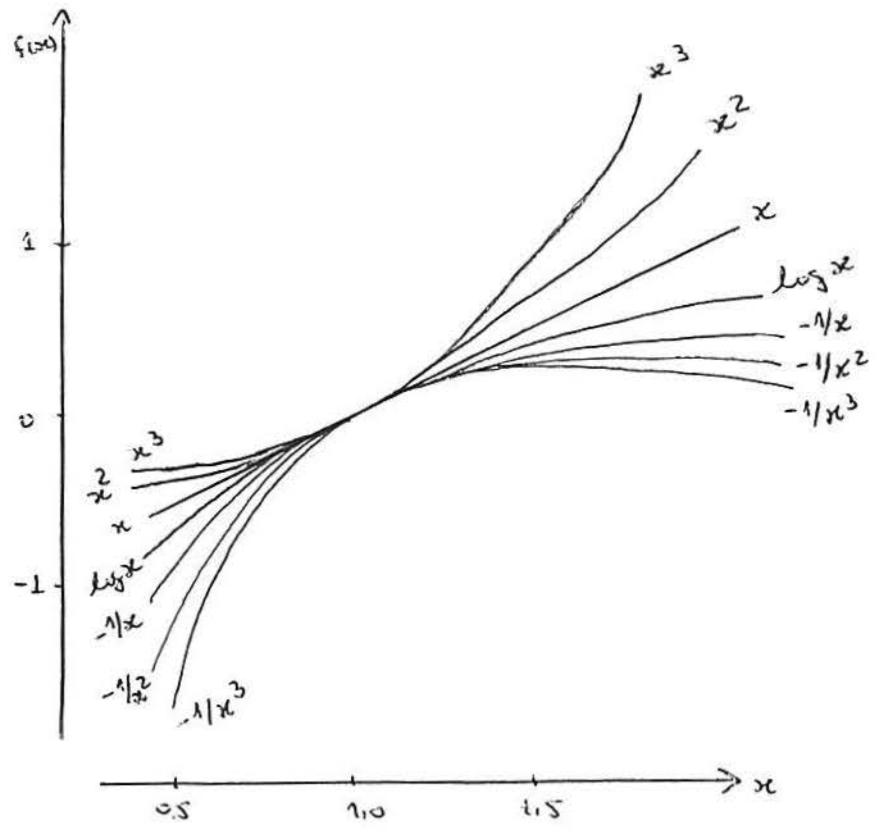
O que se busca - quase sempre - então, é uma transformação do tipo representado no quadro 19 que pelo menos diminua assimetrias - do tipo que surge quando se tem uma ocorrência do gênero "lei anormal" dos grandes números. Como se vê aí, para valores grandes ocorre uma contração, quando comparada a valores pequenos.

Quadro 19 - Efeito de uma transformação sobre o comprimento de dois intervalos iguais-um de valores "grandes", outro de valores "pequenos".



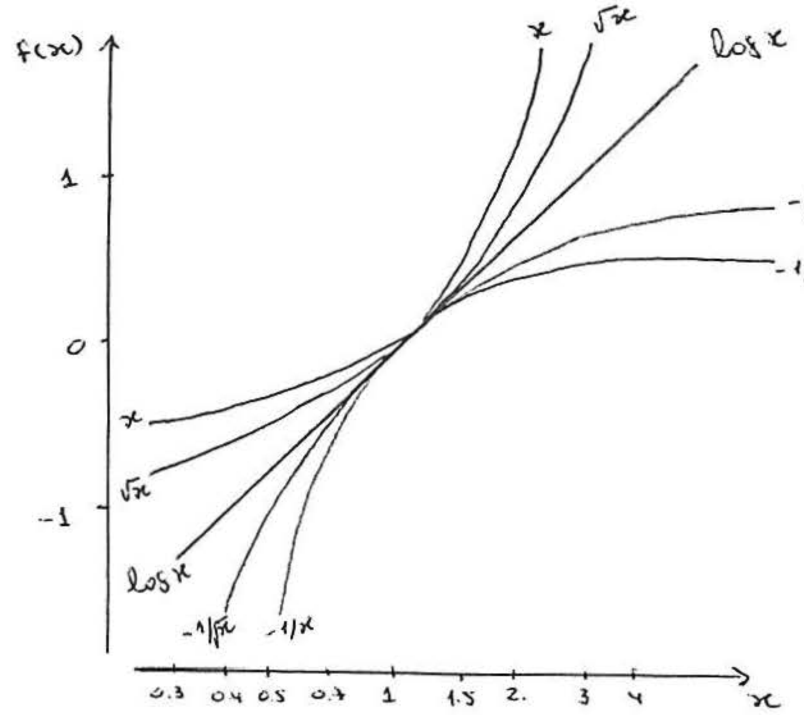
Se a assimetria fosse em sentido contrário ter-se-ia que buscar uma transformação que tivesse o efeito oposto ao da que é apresentada no quadro 19. No quadro 20 são apresentadas diversas transformações - inclusive a não-transformação - a identidade. A seguir, no quadro 21, com a escala horizontal logaritmica, são mostradas novamente algumas das mesmas transformações e algumas outras. Aí pode-se ver o papel importante da transformação logaritmica como intermediária entre $\sqrt{n}$ e $-1/\sqrt{x}$. importante ressaltar que o conjunto de transformações apresenntado é em geral suficiente para resolver os problemas encontrados na prática - principalmente devido a que, com a imprecisão nos valores e com os limites possíveis e razoáveis de percepção - outras trasnsformações intermediárias são um refinamento complicado, desnecessário e injustificado.

<u>Quadro 20</u> - Algumas transformações - com escala e origem alteradas para melhor visualização.



Função:	No gráfico:
x^3	$-1/3 + x^3/3$
x^2	$-1/2 + x\ /2^2$
x	$-1 + x$
log x	$2{,}303 \log_{10} x$
$-1/x$	$1 - 1/x$
$-1/x^2$	$1/2 - 1/2(1/x^2)$
$-1/x^3$	$1/3 - 1/3(1/x^3)$

Quadro 21 - O papel central da transformação logaritmica. Pode-se dizer que equivale potência de zero no gráfico abaixo: x^{1}, $x^{1/2}$, log, $x^{-1/2}$, x^{-1}.



Função:	No gráfico:
x	$-1 + x$
$\sqrt{x}$	$-2 + 2\sqrt{x}$
$\log x$	$2{,}303 \log x$
$-1/\sqrt{x}$	$2 + 2\,(-1/\sqrt{x})$
$-1/x$	$1 + (-1/x)$

A razão dos negativos nas transformações $-1\sqrt{x}$, $-1/x$, $-1/x^{2}$ e $-1/x^{3}$é de que assim se conserva a ordem dos valores com que se está trabalhando. Todas as transformações apresentadas são quando os valores são todos positivos - essencialmente, pelo menos as mais importantes ($\sqrt{}$, log, $-1/\sqrt{}$). O que fazer quando há valores negativos e positivos já é mais complexo e foge ao objetivo destas notas.Para resolver esses problemas pode-se consultar Tukey , capítulo 3.

VII- AULAS DE LABORATÓRIO

AULA DE LABORATÓRIO Nº 1 - MAT271 - ESTATÍSTICA DESCRITIVA

PROF. DINARA

OBJETIVO: Utilizar o programa MINITAB.

C:\ MINITAB > MINITAB

MTB > READ C1 - C2 O programa apresenta duas colunas para entrada dos dados (uma para cada variável).

DATA >

DATA >

DATA >

DATA > END Indica término dos dados.

MTB > PRINT C1 - C2 Mostra os dados na tela.

MTB > SAVE 'A:<nome arquivo>' Salva o arquivo no drive A.

MTB > OUTFILE='PRINTER' Imprime os resultados e os comandos.

MTB > STOP Sai do programa ou termina arquivo.

MTB > RETRIVE 'A:<nome arquivo>' Chama arquivo do drive A.

OUTROS COMANDOS ÚTEIS:

MTB > LET C3=C1+C2 Soma as colunas C1 e C2 por linha e coloca o resultado na coluna C3. Este comando permite fazer qualquer operação aritmética, tais como:

LET C4=logten(C1) Coloca log decimal de C1 em C4.

LET C5=SQRT(16) Coloca raiz quadrada de 16 na C5.

LET C6=5.4 Coloca a constante 5.4 em C6.

LET C7=antilog(C1) Coloca antilog de C1 em C7.

LET C8=SUM(C1) Soma C1 e coloca em C8.

LET C9=C8/10 Divide C8 por 10 e coloca em C9.

LET C10=C1*C2 Multiplica C1 por C2 e coloca em C10

MTB > ERASE Apaga dado.

MTB > PARSOM C1 Acumula os dados da coluna C1.

MTB > COUT C1 Conta o número de dados de C1.

MTB > MEAN C1 Média da coluna C1

MTB > SUM C1K1 Soma a coluna C1 e coloca o resultado na coluna constante K1.

MTB > NAMES C1 'SEX' Atribui o nome SEX à coluna C1.

MTB > MIN C1 Mínimo da coluna C1

MTB > STDEV C1 Desvio Padrão da coluna C1

MTB > HIST C1 Histograma da coluna C1

MTB > PLOT C2(x) vs C1(y) Diagrama de dispersão (gráfico)

MTB > SORT C1 Ordena os dados de C1

EXERCÍCIO: Com os dados que seguem, referentes à produção de milho em ton/ha, calcule a média geométrica e a média harmônica utilizando o MINITAB:

80.7	81.6	83.5	83.8	87.5
90.6	91.8	92.4	95.6	100.4

SOLUÇÃO DA AULA DE LABORATÓRIO Nº 1.

```
tandard Version *** Storage Available: 16179
CT. 9, 1988

TB > read c1
ATA> 80.7
ATA> 81.6
ATA> 83.5
ATA> 83.8
ATA> 87.5
ATA> 90.6
ATA> 91.8
ATA> 92.4
ATA> 95.6
ATA> 100.4
ATA> end
    10 ROWS READ
TB > mean c1 c2
  MEAN      =        88.790
TB > logten c1 c3
TB > sum c3 c4
  SUM       =        19.473
TB > let c5=c4/10
TB > antilog c5 c6
   ANSWER =           88.5796
TB >

TB >
TB >
TB >
TB >
TB >
TB >
TB >
TB >
TB >
TB > print c1-c6
```

ROW	C1	C2	C3	C4	C5	C6
1	80.7	88.79	1.90687	19.4733	1.94733	88.5796
2	81.6		1.91169			
3	83.5		1.92169			
4	83.8		1.92324			
5	87.5		1.94201			
6	90.6		1.95713			
7	91.8		1.96284			
8	92.4		1.96567			
9	95.6		1.98046			
10	100.4		2.00173			

```
TB >
```

AULA DE LABORATÓRIO Nº 2 - MAT 271 - ESTATÍSTICA DESCRITIVA
PROF. DINARA

OBJETIVO: Conhecer outras formas de acessar ao MINITAB.

Para guardar uma série de comandos (programa a ser executado) do MINITAB, podemos proceder de duas maneiras:

(1) Se estivermos dentro do MINITAB, criamos o arquivo através do comando:

```
MTB > STORE 'A:<arquivo>'
```

segue o conjunto de comandos do programa, por exemplo:

```
MTB > LET C3=C1*C2
MTB > LET C4=SUM(C3)/SUM(C2)
MTB > END
```

Depois entra-se com os dados:

```
MTB > READ C1-C2
          C1        C2
DATA >
DATA >
DATA > END
MTB > EXECUTE 'A:<arquivo>'
```

Para trocar os valores de uma coluna, fazemos:

```
MTB > SET C2
```

e digita-se os novos valores para C2.

```
MTB > EXECUTE 'A:<arquivo>'
```

e o programa roda com C1 antigo e os valores atualizados para C2 .

O processo pode ser repetido quantas vezes se fizer necessário.

(2) Criar um arquivo Não-Documento com os comandos do programa, em qualquer editor de texto, seguindo a seguinte estrutura:

conjunto de comandos

END

Entrar no MINITAB com os dados e chamar o programa:

```
MTB > READ C1-C2
          C1    C2
DATA >
DATA >
DATA > END
MTB > EXECUTE 'A:<arquivo>'
```

Para trocar os valores das colunas, procede-se analogamente à forma acima explicada.

OUTROS COMANDOS IMPORTANTES:

1. Estatísticas:

MTB >DESCRIBE Fornece várias estatísticas descritivas para as colunas solicitadas: N, MEAN, MEDIAN, STDEV, SEMEAN, MAX, MIN, Q3, Q1.

MTB > CORRELATION C1 C2 Coeficiente de correlação de Pearson entre os dados das colunas C1 e C2

Limitação: Entrando com os dados organizados em Distribuição de Frequência, não pode pedir MEAN ou qualquer outra estatística diretamente, pois o MINITAB só trabalha com colunas. Nesse caso, é necessário calcular passo a passo.

2. Histograma

MTB > HISTOGRAM C1 C2 C3 Histograma das colunas.

MTB > HISTOGRAM C1 2 5 Histograma de C1 com 1º ponto médio 2 e intervalo de tamanho 5.

MTB > PLOT y C1 x C2   Gráfico em duas dimensões com C1 em y e C2 em x. Utiliza símbolo * para 1 ponto e + para mais de 1 ponto. Tem várias opcões para troca de escala (ver manual).

3. Regressão:

MTB > REGRESS C2 on 1 PREDICTOR C1   Variável dependente y em C2 e k variáveis independentes em C2, C3, ...

4. Geração de dados aleatórios:

MTB > URANDON 20 OBSERVATIONS INTO C1   Gera 20 observações segundo uma Distribuição Uniforme entre 0 e 1 e coloca em C1.

MTB > NRANDON k OBSERVATIONS com μ=k e σ=k coloca em C Distribuição Normal.
Ex: NRANDON 20 OBSERVATIONS MU=35 SIGMA=2

MTB > RANDON k OBSERVATIONS INTO C1

5. Análise Exploratória de Dados (EDA):

MTB > BOXPLOTS C1 Faz o diagrama de caixa para C1

EXERCÍCIO: Resolva oexercício 12 da página 36, utilizando o MINITAB.

AULA DE LABORATÓRIO Nº 3 - MAT 271 - ESTATÍSTICA DESCRITIVA
Prof. Dinara - 92/2

Utilizando o programa MINITAB resolva os exercícios emitindo uma listagem dos comandos e resultados (outputs).

EXERCÍCIO 1: Utilizando o Cadastro das 80 Empresas e a Tabela de Números Aleatórios (ver Anexo), selecione uma amostra de 20 Empresas, anote os valores das variáveis: número de empregados, lucro anual e empresas exportadoras. Codifique por 1 as empresas exportadoras e por 0 as demais.

1) Calcular para as variáveis quantitativas: a) média aritmética; b) variância absoluta; c) desvio padrão; d) coeficiente de variação.

Interpretar os resultados e indicar relativamente a qual das duas variáveis as empresas são mais homogêneas.

2) Calcular a proporção de empresas exportadoras.

3) Padronizar as variáveis quantitativas. Calcular a média e o desvio padrão das variáveis padronizadas. O que podes concluir ?

4) Descobrir como ordenar os dados utilizando o MINITAB e calcular para as variáveis quantitativas: a) mediana; b)primeiro e segundo quartis; c)diferença interquartil; d)desvio mediano; e) desvio quartílico reduzido; f) coeficiente de variação de Thorndike; g)coeficiente quartílico de variação

Interpretar os resultados comparando as duas variáveis sempre que possível.

EXERCÍCIO 2: Considere os dados do exercício 1 da lista 2.3, relativamente ao efeito de drogas sobre a capacidade perceptual-motora.

1) Calcular, para cada grupo: a) média aritmética; b) variância absoluta.

2) Descobrir como agrupar os dois grupos num único, utilizando o MINITAB e calcular a média aritmética e a variância absoluta das 12 observações consideradas conjuntamente

3) Estudar o ítem da apostila DESVIO PADRÃO DE DISTRIBUIÇÃO COMBINADA (dentro de 2.1.5) e aplicar a fórmula prática para se obter a variância de várias séries neste exercício.

EXERCÍCIO 3: Estudar o ítem da apostila DECOMPOSIÇÃO DO DESVIO PADRÃO DE VÁRIAS SÉRIES (dentro de 2.1.5) e calcular: a) variância total; b) variância dentro; c) variância entre; para os dados do Exercício anterior, acrescentando duas observações na Droga 1: 5 9

Interpretar os resultados.

AULA DE LABORATÓRIO Nº 4 - MAT 271 - ESTATÍSTICA DESCRITIVA
Prof. Dinara - 92/2

Utilizando o programa MINITAB resolva o exercício a seguir, emitindo uma listagem dos programas e resultados.

Dadas as distribuições de frequência,

classes	f_{1i}	f_{2i}	f_{3i}	f_{4i}	f_{5i}	f_{6i}	f_{7i}	f_{8i}	f_{9i}	f_{10i}
0 |— 2	20	20	40	80	80	20	120	8	8	120
2 |— 4	40	60	40	32	40	20	40	12	32	32
4 |— 6	80	40	40	4	40	40	20	20	120	32
6 |— 8	40	60	40	32	20	40	12	40	32	8
8 |— 10	20	20	40	80	20	80	8	120	8	8
Σ	200	200	200	200	200	200	200	200	200	200

determinar e, sempre que possível, interpretar, comparando as distribuições quanto à:

a) momento natural de 1ª, 2ª , 3ª e 4ª ordens;

b) momento central de 2ª, 3ª e 4ª ordens, a partir dos momentos naturais;

c) média aritmética; d) variância absoluta; e)desvio padrão;

f) coeficiente de variação;

g) coeficiente de assimetria de Pearson;

h) coeficiente momento de curtose;

i) moda; j) mediana; k) primeiro e terceiro quartis;

l) primeiro e segundo coeficientes de assimetria de Pearson;

m) coeficiente quartil de assimetria; n) percentil 10 e 90;

o) coeficiente de assimetria entre os percentis 10 e 90;

p) desvio mediano; q) coeficiente percentílico de curtose.

r) coeficiente momento de assimetria.

Faça um quadro-resumo dos resultados.

distribuições / medidas	1	2	3	4	5	6	7	8	9	10
μ'_1										
μ'_2										
μ'_3										
μ'_4										
μ_2										
μ_3										
μ_4										
μ										
σ^2										
σ										
γ										
α_3										
α_4										
β_1										
β_2										
M_o										
M_e										
Q_1										
Q_3										
C_{10}										
C_{90}										
e_1										
e_Q										
e_C										
e_{M1}										
K										
Interpretação										
Gráfico (esboço)										

OLUÇÃO DA AULA DE LABORATÓRIO Nº 4.

```
TB > EXECUTE 'B:DROGA'
TB > let c3=c1*c2
TB > sum c3 c4
  SUM      =       1000.0
TB > let c5=c1*c3
TB > sum c5 c6
  SUM      =       5960.0
TB > let c7=c1*c5
TB > sum c7 c8
  SUM      =        39400
TB > let c9=c1*c7
TB > sum c9 c10
  SUM      =       280520
TB > let c11=c4/200
TB > let c12=c6/200
TB > let c13=c8/200
TB > let c14=c10/200
TB > let c15=c12-(c11**2)
TB > let c16=c13-3*c11*c12+2*(c11**3)
TB > let c17=c14-4*c11*c13+6*(c11**2)*c12-3*(c11**4)
TB > let c18=c4/200
TB > let c19=c15*1
TB > sqrt c19 c20
   ANSWER =          2.1909
TB > let c21=c20/c18
TB > let c22=(c16**2)/(c15**3)
TB > let c23=c17/(c20**4)
TB > end
TB > PRINT C1-C23
```

ROW	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	20	20	1000	20	5960	20	39400	20	280520
2	3	40	120		360		1080		3240	
3	5	80	400		2000		10000		50000	
4	7	40	280		1960		13720		96040	
5	9	20	180		1620		14580		131220	

ROW	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
1	5	29.8	197	1402.6	4.8	0	57.6001	5	4.8	2.19089

ROW	C21	C22	C23
1	0.438178	0	2.50001

```
MTB > PRINT C1-C23

 ROW   C1    C2    C3    C4    C5    C6    C7    C8    C9    C10

   1    1    20    20  1000    20  6120    20  41800    20  305160
   2    3    60   180         540        1620         4860
   3    5    40   200        1000        5000        25000
   4    7    60   420        2940       20580       144060
   5    9    20   180        1620       14580       131220

 ROW  C11   C12   C13     C14   C15  C16   C17  C18   C19      C20

   1    5  30.6   209  1525.8   5.6    0  60.8    5   5.6  2.36643

 ROW        C21   C22       C23

   1   0.473286     0   1.93878

MTB > NOOUTFILE




MTB > PRINT C1-C23

 ROW   C1    C2    C3    C4    C5    C6    C7    C8    C9    C10

   1    1    40    40  1000    40  6600    40  49000    40  386760
   2    3    40   120         360        1080         3240
   3    5    40   200        1000        5000        25000
   4    7    40   280        1960       13720        96040
   5    9    40   360        3240       29160       262440

 ROW  C11   C12   C13     C14   C15  C16   C17  C18   C19      C20

   1    5    33   245  1933.8     8    0  108.8    5     8  2.82843

 ROW        C21   C22   C23

   1   0.565685     0   1.7

MTB > NOOTFILE
* ERROR * NAME NOT FOUND IN DICTIONARY

MTB > NOOUTFILE
```

```
'B > PRINT C1-C23

:OW    C1     C2      C3      C4      C5      C6       C7       C8        C9       C10

 1      1     80      80    1140      80    8516       80    70740        80    606884
 2      3     32      96             288              864               2592
 3      5      4      20             100              500               2500
 4      7     32     224            1568            10976              76832
 5      9     80     720            6480            58320             524880

 OW     C11      C12      C13       C14      C15        C16        C17     C18

 1      5.7    42.58    353.7   3034.42    10.09   -4.03201    103.805     5.7

 OW      C19        C20         C21          C22         C23

 1     10.09    3.17648    0.557277    0.0158259     1.01962

'B > NOOUTFILE
```

```
'B > PRINT C1-C23

 OW    C1     C2      C3      C4      C5      C6       C7       C8        C9       C10

 1      1     80      80     720      80    4040       80    27600        80    207560
 2      3     40     120             360             1080               3240
 3      5     40     200            1000             5000              25000
 4      7     20     140             980             6860              48020
 5      9     20     180            1620            14580             131220

 OW     C11     C12     C13      C14     C15      C16       C17     C18     C19

 1      3.6    20.2     138   1037.8    7.24   13.152   117.467     3.6    7.24

 OW         C20         C21          C22         C23

 1    2.69073    0.747424    0.455793     2.24099

 B > NOOUTFILE
```

```
B > RPINT C1-C23
ERROR * NAME NOT FOUND IN DICTIONARY


B > PRINT C1-C23

OW    C1     C2      C3      C4      C5      C6      C7      C8       C9      C10

 1     1     20      20    1280      20    9640      20   77600       20   647560
 2     3     20      60             180             540             1620
 3     5     40     200            1000            5000            25000
 4     7     40     280            1960           13720            96040
 5     9     80     720            6480           58320           524880

OW      C11      C12      C13       C14      C15        C16        C17      C18     C19

 1      6.4     48.2      388    3237.8     7.24   -13.1520    117.468      6.4    7.24

OW          C20         C21          C22        C23

 1      2.69072    0.420426     0.455796      2.241

B > NOOUTFILE



B > PRINT C1-C23

OW    C1      C2      C3      C4      C5      C6      C7      C8       C9      C10

 1     1     120     120     496     120    2216     120   13648      120    97160
 2     3      40     120             360            1080             3240
 3     5      20     100             500            2500            12500
 4     7      12      84             588            4116            28812
 5     9       8      72             648            5832            52488

OW     C11       C12       C13       C14        C15         C16         C17      C18

 1     2.48     11.08     68.24     485.8     4.9296     16.3108     104.256     2.48

OW        C19         C20         C21         C22         C23

 1     4.9296     2.22027     0.89527     2.22083     4.29018

B > NOOUTFILE
```

```
'B > PRINT C1-C23

'OW   C1      C2      C3      C4      C5      C6      C7      C8       C9

 1    1       8       8     1504      8    12296      8   104032        8
 2    3      12      36              108             324               972
 3    5      20     100              500            2500             12500
 4    7      40     280             1960           13720             96040
 5    9     120    1080             9720           87480            787320

'OW       C10     C11      C12       C13      C14      C15       C16       C17

 1    896840     7.52    61.48    520.16   4484.2   4.9296  -16.3108   104.255

OW     C18      C19       C20        C21       C22       C23

 1    7.52   4.9296   2.22027   0.295249   2.22083   4.29016

'B > NOOUTFILE
```

```
B > PRINT C1-C23

OW    C1      C2      C3      C4      C5      C6      C7      C8      C9      C10

 1    1       8       8     1000      8     5512      8    32680      8   206920
 2    3      32      96              288             864            2592
 3    5     120     600             3000           15000           75000
 4    7      32     224             1568           10976           76832
 5    9       8      72              648            5832           52488

OW    C11     C12      C13      C14     C15    C16      C17    C18    C19    C20

 1      5   27.56    163.4   1034.6    2.56      0  25.6001      5   2.56    1.6

OW     C21    C22      C23

 1    0.32      0  3.90627

'B > NOOUTFILE
```

```
B > PRINT C1-C23

OW    C1     C2     C3     C4     C5     C6     C7     C8       C9     C10

 1     1    120    120    504    120   2248    120   13560     120   94408
 2     3     32     96           288           864            2592
 3     5     32    160           800          4000           20000
 4     7      8     56           392          2744           19208
 5     9      8     72           648          5832           52488

OW     C11      C12      C13      C14       C15       C16       C17      C18

 1    2.52    11.24     67.8   472.04    4.8896   14.8316   95.9043     2.52

OW      C19        C20         C21        C22        C23

 1   4.8896    2.21124    0.677478    1.88173    4.01136

B > STOP
* Minitab Release 6.1.1 *** Minitab, Inc. ***
orage available 16179
```

AULA DE LABORATÓRIO Nº 5 - MAT 271 - ESTATÍSTICA DESCRITIVA
Prof. Dinara - 92/2

Utilizando o software estatístico MIMITAB, resolver os exercícios.

1. Com os dados do exercício 5.12. do Capítulo V, determine:
a. coeficiente de correlação.
b. trace o diagrama de dispersão
c. ajuste a reta de mínimos quadrados.

2. Distribuição dos Alunos da Turma W, segundo a Altura, em cm, e o Peso, em kg

Altura\ Peso	45¦— 50	50¦— 55	55¦— 60	Σ
150¦—155	10	15	25	50
155¦—160	10	20	10	40
160¦—165	5	15	10	30
165¦—170	0	0	5	5
Σ	25	50	50	125

Determine:
a. coeficiente de correlação.
b. trace o diagrama de dispersão
c. ajuste a reta de mínimos quadrados.

OBSERVAÇÃO: Entregue a listagem dos PROGRAMAS e dos RESULTADOS e um RELATÓRIO interpretando os resultados obtidos e justificando a adequacidade ou não do ajuste feito aos dados. (pode ser feito em grupo de, no máximo 5 elementos)

SOLUÇÃO DA AULA DE LABORATÓRIO Nº 5.

```
MTB > correlation c1 c2

Correlation of C1 and C2 = 0.928

MTB > plot c2 c1

         -                                                    *
         -                                               *
    0.140+
         -
C2       -
         -
         -
    0.105+
         -                                     *
         -                      *    *
         -                      *    *
         -                 2    2         *
    0.070+                 *    *
         -
         -       *
         -
         -
          -------+---------+---------+---------+---------+---------+C1
               0.060     0.080     0.100     0.120     0.140     0.160

MTB > nooutfile


MTB > regress c1 on 1 predictor c2

The regression equation is
C1 = 0.0255 + 0.821 C2

Predictor        Coef        Stdev     t-ratio        p
Constant     0.025460     0.008437        3.02    0.010
C2            0.82059      0.09161        8.96    0.000

s = 0.009170     R-sq = 86.1%     R-sq(adj) = 85.0%

Analysis of Variance

SOURCE        DF          SS           MS          F        p
Regression     1   0.0067469    0.0067469      80.24    0.000
Error         13   0.0010931    0.0000841
Total         14   0.0078400

Unusual Observations
Obs.      C2        C1        Fit  Stdev.Fit   Residual   St.Resid
  1    0.154   0.15000    0.15183    0.00646   -0.00183     -0.28 X
  6    0.078   0.11000    0.08947    0.00255    0.02053      2.33R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

```
TB > correlation c1 c2

orrelation of C1 and C2 = 0.024

TB > plot c1 c2

1         -                                                             5
          -
          -
   165.0+
          -
          -       5                        +                            +
          -
          -
   160.0+
          -
          -       +                     2  +                            +
          -
          -
   155.0+
          -
          -       +                        +                            +
          -
           --------+---------+---------+---------+---------+---------+C2
                 48.0      50.0      52.0      54.0      56.0      58.0

B > regress c2 on 1 predictor c1

e regression equation is
 = 50.3 + 0.0200 C1

edictor        Coef        Stdev     t-ratio        p
nstant        50.35        11.91        4.23    0.000
            0.02000      0.07578        0.26    0.792

= 3.774        R-sq = 0.1%       R-sq(adj) = 0.0%

alysis of Variance

URCE         DF          SS          MS         F        p
gression      1        0.99        0.99      0.07    0.792
ror         123     1751.50       14.24
al          124     1752.49

usual Observations
.          C1         C2         Fit  Stdev.Fit   Residual    St.Resid
          168     57.500      53.700      0.857      3.800      1.03 X
          168     57.500      53.700      0.857      3.800      1.03 X
          168     57.500      53.700      0.857      3.800      1.03 X




          168     57.500      53.700      0.857      3.800      1.03 X
          168     57.500      53.700      0.857      3.800      1.03 X

enotes an obs. whose X value gives it large influence.
```

VIII- RESPOSTAS DE ALGUNS EXERCÍCIOS

pág. 2:

1. podemos; permite; podemos; permite; podemos; permite; podemos; permite; escala de razão

2. 2.1. Escala intervalar pois não tem zero absoluto.
   2.2. Ordinal
   2.3. Nominal

3. 3.1. Escala de razão
   3.2. Nominal
   3.3. Ordinal

4. Nominal

5. Ordinal

6. Não

7. Escala de razão

8. de razão

pág. 34:

5.

	A	B	classificação
μ	6,375	6,375	B = A
M_h	6,1426	6,3488	B > A
M_g	6,2579	6,3618	B > A
M_q	6,4903	6,3885	A > B

6. a) Não; b) $151 \leq Y \leq 185$; c) 167

7. $N = 9$; $\mu = 72$

11. [150 ; 162,5]
    [162,5 ; 168,03]
    [168,03 ; 173,15]
    [173,15 ; 185]

12. a) a.1) 120,53 a.2) 53 a.3) 25
    b) $M_o < M_e < \mu$ $\therefore$ Assimetria positiva
    c) M_o ou M_e

15. a) 24,34 ; b) 25,33 ; c) 19; d) 5130; e) 24,4; f) $Q_2 = M_e$

pg.56

1.

	Droga 1	Droga 2
μ	7,6	5,0
σ	1,0198	1,0690
γ	0,1342	0,2130

Droga 1 é mais homogênea.

2. $Z_{CV} = 1,047$; $Z_{AN} = 2,05$

Não, levando em conta o grupo, o resultado foi superior em AN.

3. $\mu_{Total} = 6,206$; $\sigma_{Total} = 2,3133$

7. $X_1 = 13$ e $X_2 = 7$

8. $\mu_Y = 64$; $\sigma_Y = 7,5$

11.

	A	B	C
μ	20	10	8
σ	5	3	4
γ	0,25	0,3	0,5

C é a mais homogênea.

16.

	A	B
μ	20	20
σ	0,2646	0,2582
γ	0,01323	0,0129

B deve ser preferida.

18.

	A	B
μ	7,54	5,68
σ	0,9962	0,6615
γ	0,1321	0,1165

B é mais homogênea do que A.

22. a) Não houve perda.
b) $\mu = 0,2498$; $\sigma = 0,0048$; $\gamma = 0,01938$

pág. 93.

5.1. a) 66,91% ; b) 33,09% ; c) 66,91% ; d) 0,3309%

5.3. a) Distribuição dos pais segundo Método de Educação

Método de Educação (X)	fi
Permissivo	30
Moderado	28
Autoritário	31
Total	89

Distribuição dos pais segundo Orientação Política

Orientação Política	fi
Conservador	32
Moderado	30
Liberal	37
Total	89

b) Distribuição dos pais com Método de Educação Permissiva segundo a Orientação Política (Y/X_1)

Orientação Política	f_{1j}	%
Conservador	7	23
Moderado	9	30
Liberal	14	46
Total	30	100

Distribuição dos pais com Método de Educação Moderada segundo a Orientação Política (Y/X_2)

Orientação Política	f_{2j}	%
Conservador	10	35
Moderado	10	35
Liberal	8	28
Total	28	100

Distribuição dos pais com Método de Educação Autoritária segundo a Orientação Política (Y/X_3)

Orientação Política	f_{3j}	%
Conservador	15	48
Moderado	11	35
Liberal	5	17
Total	31	100

Distribuição dos pais com Orientação Política Conservadora segundo o Método de Educação (X/Y_1)

Método de Educação	f_{i1}	%
Permissivo	7	21
Moderado	10	31
Autoritário	15	46
Total	32	100

Distribuição dos pais com Orientação Política Moderada segundo o Método de Educação (X/Y_2)

Método de Educação	f_{i2}	%
Permissivo	9	30
Moderado	10	33
Autoritário	11	36
Total	30	100

Distribuição dos pais com Orientação Política Liberal segundo o Método de Educação (X/Y_3)

Método de Educação	f_{i3}	%
Permissivo	14	51
Moderado	8	29
Autoritário	5	18
Total	27	100

c) Os pais de Educação Permissiva tendem a ser mais liberais politicamente.
Os pais de Educação Moderada não diferem muito quanto à Orientação Política.
Os pais de Educação autoritátia são mais conservadores.
Portanto, observa-se realmente uma associação entre Método de Educação e Orientação Política.

5.5.

	Prep. Formal	Sem Prep. Formal	
Aprov.	22	10	32
Reprov.	8	18	26
	30	28	58

Distribuição dos alunos com Prep. Formal segundo o Aproveitamento ($X/Y=1$)

Aproveitamento	f_{r1j}
Aprovado	22/30 = 0,73
Reprovado	8/30 = 0,27

Distribuição dos alunos sem Prep. Formal
segundo o Aproveitamento (X/Y=2)

Aproveitamento	f_{r2j}
Aprovado	10/28 = 0,35
Reprovado	18/28 = 0,64

Os alunos com Preparação formal obtiveram melhor aproveitamento do que os outros.

5.7. b) $\rho = 0,61$
c) $\beta = 0,7689$ $\alpha = -0,17$

IX- REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, M.I.F. Introdução à análise exploratória de dados. Seminário apresentado na ESALQ/USP, Piracicaba, 1987.

ANDREWS,H.; SNEE,R. & SARNER,M. (1980). Graphical display of means. **The American Statistician, 34**, 4, 195-199.

BUSSAB, W. & MORETTIN, P. **Métodos quantitativos para economistas e administradores.** Atual Editora, 1981.

COCHRAN, W. **Técnicas de Amostragem.** Ed. Fundo de Cultura, 1963.

DACHS, J.N.W. **Análise de Dados e Regressão.** Departamento de Estatística, IMECC/UNICAMP, julho 1978.

DUNN,R, (1988). Framed retangle charts or statistical maps with shading - An experiment in graphical perception. **The American Statistician, 42**, 2, 123-129.

DUNN,R. (1989) . A dynamic approach to two-variable color mapping. **The American Statistician, 43**, 4, 245-252.

ENDO, S.K. **Números Índices.** Atual Editora, 1986.

FERNANDEZ,D.W.X. **Números Índices.** Cadernos de Matemática e Estatística, Série B, nº 11, 29 p., UFRGS, outubro 1992.

FERNANDEZ, D.W.X. **Estatística Descritiva I.** Cadernos de Matemática e Estatística, Série B, nº 22, 99 p., UFRGS, janeiro 1994.

FRIGGE, M; HOAGLIN,D. & IGHEWICZ,b. (1989). Some implementations of the boxplot. **The American Statistician,43**, 1, 50-54.

GENTLEMAN,J. (1977). It's all a plot (using interactive computer graphics in teaching statistics). **The American Statistician, 31**, 4, 166-175.

HEFFERNAN,P. (1988). New measures of spread and a simpler formula for the normal distribution. **The American Statistician, 42**, 2, 100-102.

IEMMA,A.F. **Estatística Descritiva.** *φσρ* Publicações, 1992.

KAIGH,W & DRISCOO,M. (1987). Numerical and graphical data summary using O-statistics.**The American Statistician, 41**, 1, 25-32.

KAIGH,W.D. (1989). A category representation paradox. **The American Statistician,43,** 92-94.

KAZMIER, L.J. **Estatística Aplicada à Economia e Administração.** McGraw-Hill, 1982.

MCNEIL, D.R. **Interactive Data Analysis - A Practical Primer.** John Wiley and Sons, Inc., New York, 1977.

MAGE,D. (1982). An objective graphical method for testing normal distributional assumptions using probability plots. **The American Statistician, 36,** 2, 116-120.

McGILL,R.; TUKEY,J. & LARSEN,W. (1978) . Variations of box plots. **The American Statistician, 32,** 1, 12-26.

MENDENHALL, W. **Probabilidade e Estatística.** Ed. Campus,1985.

MICHEL, N.E. A teoria da preferência revelada: uma análise empírica do bem-estar de um conjunto de consumidores. **Textos para discussão/12.** Ed. Universidade, 1983.

NICK, EVA **Fundamentos de Estatística para as ciências do comportamento.** Ed. Renes, 1971.

PEREIRA, R.S. **A Estatística e suas aplicações.** Grafosul, 1979.

PEREIRA, W. & TANAKA, O.K. **Elementos de Estatística.** McGraw-Hill, 1984.

RICHARD, A. B.; WILLIAM, S.C. & ALLAN, R.W. (1987).Dynamic Graphics for Data Analysis. **Statistical Science,2,** 4, 355-395.

ROSENBAUM, P. (1989). Exploratory plots for paired data. **The American Statistician, 43,** 2, 108-109.

RUPPERT,D. (1987). What is kurtosis? - An influence fuction approach. **The American Statistician, 41**, 1, 1-5.

SIEGEL, S. **Estatística não-paramétrica.** Ed. McGraw-Hill, 1975.

SPIEGEL, M.R. **Estatística.** McGraw-Hill, 1985.

STEVENSON, W.J. **Estatística Aplicada à Administração.** Harbra, 1981.

TOLEDO, G.L. & OVALLE, I.I. **Estatística Básica.** 1985.

TRUMBO, B.E. (1981) . A Theory for coloring bivariate Statistical Maps. **The American Statistician, 35,** 220-226.

TUKEY, V.W. (1990) Data- Based Graphics:Visual display in the decades to come. **Statistical science, 5,** 3, 327-339.

TUKEY, J.W. **Exploratory Data Analysis.** Addison-Wesley Publ. Co., Reading, Mass., 1977.

WAINER,H. (1984). How to display data badly. **The American Statistician, 38,** 2, 137-147.

WAINER,H. & FRANCOLINI, C.M. (1980). An empirical inquiry concerning human understanding of two-variabel color maps. **The American Statistician, 34,** 81-93.

WAINER, H. (1990). Graphical visions from William Play Fair to John Tukey. **Statistical Science, 5,** 3, 340-346.

X- ANEXOS

ANEXO 1: FORMULÁRIOS

FORMULÁRIO 1

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \mu = \frac{\sum_{i=1}^{m} f_i \, x_i}{N} \qquad \varepsilon = \frac{h}{2\,\mu - h} \qquad M_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

$$M_g = \text{antilog} \frac{\sum_{i=1}^{N} \log x_i}{N} \qquad M_g = \sqrt[N]{\prod_{i=1}^{m} x_i^{f_i}} \qquad M_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

$$M_h = \frac{N}{\sum_{i=1}^{m} \frac{f_i}{\bar{x}_i}} \qquad M_p = \sqrt[P]{\frac{\sum^{N} x_i^p}{N}} \qquad M_q = \sqrt[2]{\frac{\sum f_i \, x_i^2}{N}}$$

$$M_e = l_i + h \left[\frac{N/2 - F_{ant}}{f_i} \right] \qquad M_o = l_i + h \left[\frac{d_1}{d_1 + d_2} \right]$$

$$M_o = l_i + h \, \frac{f_i - f_{ant}}{2\,f_i - (f_{post} + f_{ant})} \qquad M_o = l_i + h \, \frac{f_{post}}{f_{ant} + f_{post}}$$

$$M_o = 3\,M_e - 2\,\mu$$

FORMULÁRIO 2

$$H = X_{máx} - X_{min}$$

$$DM = \frac{\sum^{N} |d_i|}{N} = \frac{\sum^{N} |X_i - \mu|}{N} \qquad DM_{M_e} = \frac{\sum^{N} |X_i - M_e|}{N}$$

$$DM = \frac{\sum^{m} f_i\,|d_i|}{N} = \frac{\sum^{m} f_i\,|X_i - \mu|}{N} \qquad DM_{M_e} = \frac{\sum^{m} f_i\,|X_i - M_e|}{N}$$

$$SQ = \sum^{N} d_i^2 = \sum^{N} (x_i - \mu)^2 \qquad \sigma^2 = \frac{\sum^{N} d_i^2}{N} = \frac{\sum^{N} (x_i - \mu)^2}{N}$$

$$SQ = \sum^{m} f_i d_i^2 = \sum^{m} f_i (x_i - \mu)^2 \qquad \sigma^2 = \frac{\sum^{m} f_i d_i^2}{N} = \frac{\sum^{m} f_i (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum x_i^2}{N} - \mu^2 \qquad \varepsilon = \frac{2\,h\,\mu}{\sigma^2 - 2\,h\,\mu} \qquad \sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\sum f_i x_i^2}{N} - \mu^2 \qquad \sigma^2_{corrig} = \sigma^2 - \frac{h^2}{\sigma^2} \qquad \hat{s}^2 = \frac{n}{n-1}\,s^2$$

$$(1 - 1/k^2) \qquad DI = Q_3 - Q_1 \qquad \hat{s}^2 = \frac{N-1}{N}\,\frac{n}{n-1}\,s^2$$

$$DM = \frac{4}{5}\,\sigma \qquad DQ = \frac{Q_3 - Q_1}{2} \qquad z = \frac{x - \mu}{\sigma} \qquad \gamma^2 = \frac{\sigma^2}{\mu^2}$$

$$DQ_r = \frac{Q_3 - Q_1}{2\,M_e} \qquad CV_T = \frac{\sigma}{M_e} \qquad CV_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad \gamma = \frac{\sigma}{\mu}$$

$$\mu_G = \frac{\sum_{i=1}^{k} N_i\,\mu_i}{\sum_{i=1}^{k} N_i} \qquad \sigma_G^2 = \frac{\sum_{i=1}^{k} N_i\,\sigma_i^2 + \sum_{i=1}^{k} N_i\,\mu_i^2}{\sum_{i=1}^{k} N_i} - \mu_G^2$$

$$\sum_{ij}^{kN} x_{ij}^2 - \frac{\left(\sum^{kN} x_{ij}\right)^2}{k\,N} = \left[\sum_{ij}^{kN} x_{ij}^2 - \sum_{j}^{k} \frac{\left(\sum^{N} x_{ij}\right)^2}{N}\right] + \left[\sum_{j}^{k} \frac{\left(\sum^{N} x_{ij}\right)^2}{N} - \frac{\left(\sum^{kN} x_{ij}\right)^2}{k\,N}\right]$$

FORMULÁRIO 3

$$\mu'_r = \frac{\sum_{i=2}^{N} x_i^r}{N} \qquad \mu_r = \frac{\sum_{i=1}^{N} (x_i - \mu)^r}{N} \qquad \mu_r = \frac{\sum_{i=1}^{N} (x_i - x_0)^r}{N}$$

$$\mu'_r = \frac{\sum_{i=2}^{m} f_i x_i^r}{N} \qquad \mu_r = \frac{\sum_{i=1}^{m} f_i (x_i - \mu)^r}{N} \qquad \mu_r = \frac{\sum_{i=1}^{m} f_i (x_i - x_0)^r}{N}$$

$$\sigma^2 = \mu'_2 - (\mu'_1)^2 \qquad \mu_3 = \mu'_3 - 3\,\mu'_1\,\mu'_2 + 2\,(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\,\mu'_1\,\mu'_3 + 6\,(\mu'_1)^2\,\mu'_2 - 3\,(\mu'_1)^4$$

$$\mu_2 = {}_{x_0}\mu_2 - ({}_{x_0}\mu_1)^2$$

$$\mu_3 = {}_{x_0}\mu_3 - 3\,({}_{x_0}\mu_1)\,({}_{x_0}\mu_2) + 2\,({}_{x_0}\mu_1)^3$$

$$\mu_4 = {}_{x_0}\mu_4 - 4\,({}_{x_0}\mu_1)\,({}_{x_0}\mu_3) + 6\,({}_{x_0}\mu_1)^2\,({}_{x_0}\mu_2) - 3\,({}_{x_0}\mu_1)^4$$

$$\alpha_r = \frac{\mu_r}{\sigma^r} \qquad \beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3} \qquad \beta_2 = \alpha_4 = \frac{\mu_4}{\sigma^4} \qquad C_2 = \beta_2 - 3$$

$$e_1 = \frac{\mu - M_o}{\sigma} \qquad e_2 = \frac{3\,(\mu - M_e)}{\sigma}$$

$$e_Q = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} \qquad e_C = \frac{(C_{90} - M_e) - (M_e - C_{10})}{(C_{90} - M_e) + (M_e - C_{10})}$$

$$e_{M_1} = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\,\beta_2 - 6\,\beta_1 - 9)} \qquad K = \frac{D_Q}{C_{90} - C_{10}}$$

FORMULÁRIO 4

$$f_r\,(X_i / Y_j) = \frac{f_{ij}}{f_{.j}} \qquad f_{ij} = \frac{f_{i.}\ .\ f_{.j}}{N}$$

$$\mu_{xy} = \frac{1}{N} \sum_i \sum_j x_i\, y_j\, f_{ij} \quad ; \quad \mu_x = \frac{1}{N} \sum f_{i.}\, x_i \quad ; \quad \mu_y = \frac{1}{N} \sum f_{j.}\, y_j$$

$$\mu_{xy} = \mu_x\ .\ \mu_y \qquad \mathrm{Cov}\,(X,Y) = \sigma_{xy} = \frac{1}{N} \sum_i \sum_j (x_i - \mu_x)\,(y_j - \mu_y)\, f_{ij}$$

$$\mathrm{Cov}(X,Y) = \sigma_{XY} = \mu_{xy} - \mu_x\ .\ \mu_y \qquad \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\ \sigma_y}$$

$$\rho_{xy} = \frac{N \sum_i \sum_j x_i\, y_j\, f_{ij} - \left[\sum f_{i.}\, x_i\right] \left[\sum f_{j.}\, y_j\right]}{\sqrt{\left[N \sum f_{i.}\, x_i^2 - \left(\sum f_{i.}\, x_i\right)^2\right]\left[N \sum f_{j.}\, y_j^2 - \left(\sum f_{j.}\, y_j\right)^2\right]}}$$

$$\alpha = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum x\, y)}{N \sum x^2 - (\sum x)^2}$$

$$\beta = \frac{N \sum x\, y - (\sum x)\,(\sum y)}{N \sum x^2 - (\sum x)^2}$$

$$\alpha = \overline{Y} - \beta\, \overline{X}$$

$$\varepsilon_{XY} = \sqrt{\frac{\sum (y_i - y_{est})^2}{N}} \qquad \rho^2 = \frac{\beta^2\ \sigma_x^2}{\sigma_y^2}$$

$$\varepsilon_{XY} = \sqrt{\frac{\sum y_i^2 - \alpha \sum y_i - \beta \sum x_i\, y_i}{N}}$$

ANEXO 2: TABELA DE NÚMEROS ALEATÓRIOS

07018	31172	12572	23968	55216	85366	56223	09300	94564	18172
52444	65625	97918	46794	62370	59344	20149	17596	51669	47429
72161	57299	87521	44351	99981	55008	93371	60620	66662	27036
17918	75071	91057	46829	47992	26797	64423	42379	91676	75127
13623	76165	43195	50205	75736	77473	07268	31330	07337	55901
27426	97534	89707	97453	90836	78967	00704	85734	21776	85764
96039	21338	88169	69530	53300	29895	71507	28517	77761	17244
68282	98888	25545	69406	29470	46476	54562	79373	72993	98998
54262	21477	33097	48125	92982	98382	11265	25366	06636	25349
66290	27544	72780	91384	47296	54892	59168	83951	91075	04724
53348	39044	04072	62210	01209	43999	54952	68699	31912	09317
34482	42758	40128	48436	30254	50029	19016	56837	05206	33851
99268	98715	07545	27317	52459	75366	43688	27460	65145	65429
95342	97178	10401	31615	95784	77026	33087	65961	10056	72834
38556	60373	77935	64608	28949	94764	45312	71171	15400	72182
39159	04795	51163	84475	60722	35268	05044	56420	39214	89822
41786	18169	96649	92406	42773	23672	37333	85734	99886	81200
95627	30768	30607	89023	60730	31519	53462	90489	81693	17849
98738	15548	42263	79489	85118	97073	01574	57310	59375	54417
75214	61575	27805	21930	94726	39454	19616	72239	93791	22610
73904	89123	19271	15792	72675	62175	48746	56084	54029	22296
33329	08896	94662	05781	59187	53284	28024	45421	37956	14252
66364	94799	62211	37539	80172	43269	91133	05562	82385	91760
68349	16984	86532	96186	53893	48268	82821	19526	63257	14288
19193	99621	66899	12351	72438	99839	24228	32079	53517	18558
49017	23489	19172	80439	76263	98918	59330	20121	89779	58862
76941	77008	27646	82072	28048	41589	70883	72035	81800	50296
55430	25875	26446	25738	32962	24266	26814	01194	48587	93319
33023	26895	65304	34978	43053	28951	22676	05303	39725	60054
87337	74487	83196	61939	05045	20405	69324	80823	20905	68727
81773	36773	21247	54735	68996	16937	18134	51873	10973	77090
74279	85087	94186	67793	18178	82224	17069	87880	54945	73489
34968	76028	54285	90845	35464	68076	15868	70063	26794	81386
99696	78454	21700	12301	88832	96796	59341	16136	01803	17537
55282	61051	97260	89829	69121	86547	62195	72492	33536	60137
31337	83886	72886	42598	05464	88071	92209	50728	67442	47529
94128	97990	58609	20002	76530	81981	30999	50147	93941	80754
06511	48241	49521	64568	69459	95079	42588	98590	12829	64366
69981	03469	56128	80405	97485	88251	76708	09558	86759	15065
23701	56612	86307	02364	88677	17192	23082	00728	78660	74196
09237	24607	12817	98120	30937	70666	76059	44446	94188	14060
11007	45461	24725	02877	74667	18427	45658	40044	59484	59966
60622	78444	39582	91930	97948	13221	99234	99629	22430	49247
79973	43668	19599	30021	68572	31816	63033	14597	28953	21162
71080	71367	23485	82364	30321	42982	74427	25625	74309	15855
09923	26729	74573	16583	37689	06703	21846	78329	98578	25447
63094	72826	65558	22616	33472	67515	75585	90005	19747	08865
19806	42212	41268	84923	21002	30588	40676	94961	31154	83133
17295	74244	43088	27056	86338	47331	09737	83735	84058	12382
59338	27190	99302	84020	15425	14748	42380	99376	30496	84523

*The Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates*, The Free Press, 1955.

ANEXO 3: CADASTRO DE EMPRESAS

EMPRESA	Nº EMPREGADOS	LUCRO ANUAL (Cr$)
1	60	520
2	24	680
*3	64	560
4	82	550
5	34	580
6	98	620
7	30	780
8	60	640
9	40	625
*10	24	530
*11	22	590
*12	26	610
*13	32	615
14	26	540
15	32	602
16	26	640
*17	36	650
18	24	630
19	22	550
*20	28	660
21	50	550
*22	38	580
23	60	590
24	20	665
25	26	620
26	44	580
*27	64	620
*28	40	608
*29	38	620
30	20	640
31	20	520
*32	26	640
33	24	545
34	28	680
*35	46	565
36	20	600
37	82	622
*38	46	624

EMPRESA	Nº EMPREGADOS	LUCRO ANUAL (Cr$)
39	36	582
*40	42	608
41	22	522
*42	40	666
43	38	540
*44	26	630
45	28	590
46	38	596
*47	30	599
48	26	542
49	50	608
50	24	670
51	39	640
*52	48	630
53	32	500
54	44	540
*55	42	560
56	24	620
*57	46	610
58	36	635
*59	36	642
60	48	660
61	42	520
*62	82	542
63	20	480
*64	28	600
65	56	610
*66	32	622
67	72	636
*68	32	644
69	46	682
*70	70	693
71	36	582
72	34	600
73	44	601
*74	92	604
75	76	680
*76	38	590

* EMPRESA EXPORTADORA

Publicações do Instituto de Matemática da UFRGS
Cadernos de Matemática e Estatística

Série B: Trabalho de Apoio Didático

1. Elsa Mundstock - Curso Básico Sobre Wordstar 3.45 - MAR/89

2. Jaime B. Ripoll - Introdução ao Cálculo Diferencial Via Funções de Uma Variável Real - OUT/89

3. Edmund R. Puczylowski - Dimension of Modular Lattices - JUN/90

4. Marcos Sebastiani - Geometrias Não Euclidianas - JUL/90

5. Sandra R. C. Pizzatto - Cálculo Numérico - AGO/91

6. Vera Clotilde G. Carneiro - Elementos de Calculo para Biologia - AGO/91

7. Elsa Mundstock - Iniciação ao SPSS/PC - SET/91

8. Elisa Hagg, Loiva C. de Zeni, Maria Alice Gravina e Vera Carneiro - Notas da 1ª Oficina de Matemática da UFRGS - JAN/92

9. Paulo Werlang de Oliveira, Elisabete Rambo, Suzana Lima dos Santos, Coordenação: Profª Maria Alice Gravina - A Tartaruga no Espaço Tridimensional - FEV/92

10. Silvio Possoli - Análise Multivariada - JUL/92

11. Dinara Westphalen Fernandez - Números Índices - OUT/92

12. Maria Teresinha Albanese - Coeficiente de Fidedignidade de um Instrumento de Medida - OUT/92

13. Vera Clotilde Carneiro e Sergio Claudio Ramos - Gráficos na Escola - DEZ/92

14. João Riboldi - Elementos Básicos de Estatística - JAN/93

15. Paulo W. de Oliveira e M. Alice Gravina - Logo: Manual do Usuário - MAR/93

16. Ruben Markus, Elsa C. de Mundstock, Dinara W. X. Fernandez e João Riboldi - Exercícios de Métodos Estatísticos - AGO/93

17. Loiva C. de Zeni e M. Alice Gravina - Sugestões de Atividades no Ambiente Logo para a Exploração de Conteúdos Matemáticos dos Currículos Escolares de 1º e 2º Grau - SET/93

18. João Riboldi - Delineamentos Experimentais de Campo, Parte 1 - SET/93

19. Marlusa Benedetti, Patrícia P. Gil, Shirley I. Techera, Angela Andreotti, Milene Milan, Marlise Moraes, Luciana Santos, Augustinho Zimmermann, Coordenação: Profª Maria Alice Gravina - Atividades em Geometria Usando Recortes - OUT/93

20. João Riboldi - Delineamentos Experimentais de Campo, Parte 2 - OUT/93

21. Anne C. Rutsatz, Édina R. de C. Alexandre, Gorete Losada, M. Alice Gravina, Rosamary P. Disconzi, Shirley Techera e Vera C. G. Carneiro - O Pensamento e a Linguagem da Álgebra: Tabelas, Gráficos e Equações - DEZ/93

22. Dinara W. X. Fernandez - Estatística Descritiva I - JAN/94

23. João Riboldi - Planejamento e Análise de Experimentos, Parte 1 -FEV/94

24. Dinara W. X. Fernandez - Estatística Descritiva II - AGO/94

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

NÚCLEO DE ATIVIDADES EXTRACURRICULARES

Os Cadernos de Matemática e Estatística publicam as seguintes séries:

Série A: Trabalho de Pesquisa

Série B: Trabalho de Apoio Didático

Série C: Colóquio de Matemática SBM/UFRGS

Série D: Trabalho de Graduação

Série E: Trabalho de Divulgação

Série G: Textos para Discussão

Toda correspondencia com solicitação de números publicados e demais informações deverá ser enviada para:

NAEC - NÚCLEO DE ATIVIDADES EXTRACURRICULARES

INSTITUTO DE MATEMATICA - UFRGS

AV. BENTO GONÇALVES, 9500 - PRÉDIO 43111

CEP 91509 - 900 AGRONOMIA - POA/RS

FONE: 336 92 22 OU 339 13 55 OU 228 16 33

RAMAL 6197

FAX: 336 15 12